MATH2801 Revision Session

UNSW StatSoc x MathSoc

August 12, 2020

Credit to Rui Tong's Slides

Overview

1. Probability Theory
   - Descriptive Statistics
   - Probability Theory
   - Common Distribution
   - Bivariate Distributions
   - Distribution of Sums and Averages in Random Variables
2. Statistical Inference
   - Desirable features of estimators
   - More work with Convergence
   - Confidence Intervals
   - Hypothesis tests

Categorical v.s. Quantitative

Categorical

Based off some 'category'.

E.g. Sunny v.s. Cloudy, Male v.s. Female

Quantitative

Based off some 'scale'; usually involves numbers.

E.g. Weight, Precipitation, Age lived

Course Focus - Quantitative Data

Nature of quantitative data

- Location - Where abouts is the data centered?
- Scale - To what extent is the data spread around there?
- Shape - Symmetric v.s. Skewed

Skewness of data

- *Negatively* skewed data is clustered towards the *right*.
- *Positively* skewed data is clustered towards the *left*.

Numerical summaries of Quantitative Data

Definitions

- Sample Mean

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

- Sample Variance

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

- Sample Standard Deviation

$$s = \sqrt{s^2}$$

- Sample Median
- pth Sample Quantile

Probability

Definition (2901)

A probability is a function $\mathbb{P}$ that assigns a value in $[0, 1]$ from events in the sample space Ω, in the σ-algebra (say $\mathcal{A}$).

Definition (Probability Space) (2901)

A *probability space* is the triple $(\Omega, \mathcal{A}, \mathbb{P})$ with the axioms

$$\mathbb{P}(A) \geq 0 \qquad \forall A \in \mathcal{A}$$
$$\mathbb{P}(\Omega) = 1$$
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

for mutually exclusive events $A_1, A_2, \cdots \in \mathcal{A}$

Probability

Definition (Probability Space) (2901)

A *probability space* is the triple $(\Omega, \mathcal{A}, \mathbb{P})$ with the axioms

$$\mathbb{P}(A) \geq 0 \qquad \forall A \in \mathcal{A}$$
$$\mathbb{P}(\Omega) = 1$$
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

for mutually exclusive events $A_1, A_2, \cdots \in \mathcal{A}$

Don't worry too much about them.

Complementary Event

Definition (Complement)

Given an event A, the complement A^c is essentially the event representing 'not A'

Theorem (Probability of a complement)

For any event $A \in \mathcal{A}$,

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

Conditional Probability

Definition (Conditional Probability)

Given that the event $B \in \mathcal{A}$ has occurred, the probability of $A \in \mathcal{A}$ occuring is

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

Theorem (Multiplication Law)

If $\mathbb{P}(B) \neq 0$, then the probability of A and B occurring is

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

and similarly if $\mathbb{P}(A) \neq 0$,

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

Conditional Probability

Theorem (Multiplication Law)

If $\mathbb{P}(B) \neq 0$, then the probability of A and B occurring is

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

Example (MATH1251)

A diagnostic test has 99% chance of correctly detecting if someone has a disease. If only 2% of the population have this disease, what is the probability that someone has the disease and was successfully tested for it?

$$\mathbb{P}(D \cap T) = \mathbb{P}(T \mid D)\mathbb{P}(D) = 0.99 \times 0.02$$

Independence

Definition (Independence)

Two events $A, B \in \mathcal{A}$ are independent if

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

Remark

If $\mathbb{P}(B) \neq 0$, then two events are independent iff

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

Total Probability

Theorem (Law of Total Probability)

Let the events $A_1, A_2, \ldots$ be mutually exclusive. Then

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\mathbb{P}(A_2) + \ldots \\ &= \sum_i \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).\end{aligned}$$

We can have a finite *or* infinite number of events A_i.

Bayes' Law

Theorem (Bayes' Law)

Let the events $A_1, A_2, \ldots$ be mutually exclusive. Then

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}\end{aligned}$$

Often used in conjunction with the law of total probability to obtain

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\sum_i \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}$$

Bayes' Law

Theorem (Bayes' Law)

Let the events $A_1, A_2, \ldots$ be mutually exclusive. Then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Example (MATH1251) (contd.)

99% of the people with the disease receive a positive test. 98% of those without receive a negative test. If 2% of the population have the disease, determine the probability of someone having the disease *given* they received a positive test.

Bayes' Law

Example (MATH1251) (contd.)

99% of the people with the disease receive a positive test. 98% of those without receive a negative test. If 2% of the population have the disease, determine the probability of someone having the disease *given* they received a positive test.

$$\text{We require } \mathbb{P}(D \mid T) = \frac{\mathbb{P}(T \mid D)\mathbb{P}(D)}{\mathbb{P}(T)}.$$

$$\begin{aligned}\mathbb{P}(T) &= \mathbb{P}(T \mid D)\mathbb{P}(D) + \mathbb{P}(T \mid D^c)\mathbb{P}(D^c) \\ &= \mathbb{P}(T \mid D)\mathbb{P}(D) + \big(1 - \mathbb{P}(T^c \mid D^c)\big)\mathbb{P}(D^c) \\ &= 0.99 \times 0.02 + (1 - 0.98) \times 0.98 = 0.0394\end{aligned}$$

$$\therefore \mathbb{P}(D \mid T) = \frac{0.99 \times 0.02}{0.0394} \approx 0.5025$$

Bayes' Law

A lot of people get stuck with Bayes' law, especially when used with other results. Use a tree diagram!

Discrete Random Variables

Random variable X can be defined as a function which assigns a number to each outcome in the sample space.

Definition (Discrete Random Variable)

X is a discrete random variable if it can only take countably many values.

The probability function is denoted

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

In 2801, this is also denoted $f_X(x)$ for the discrete case.

Validity of the discrete random variable

Properties of the discrete random variable

A discrete random variable must satisfy

- $\mathbb{P}(X = x) \geq 0$ for all x
- $\sum_{\text{all } x} \mathbb{P}(X = x) = 1$

Example

A discrete random satisfies $\mathbb{P}(X = 1) = \frac{1}{3}$ and $\mathbb{P}(X \neq -1, X \neq 1) = 0$. What must $\mathbb{P}(X = -1)$ equal to?

Validity of the discrete random variable

Properties of the discrete random variable

A discrete random variable must satisfy

- $\mathbb{P}(X = x) \geq 0$ for all x
- $\sum_{\text{all } x} \mathbb{P}(X = x) = 1$

Example

A discrete random satisfies $\mathbb{P}(X = 1) = \frac{1}{3}$ and $\mathbb{P}(X \neq -1, X \neq 1) = 0$. What must $\mathbb{P}(X = -1)$ equal to?

From the second property, $\mathbb{P}(X = -1) = 1 - \frac{1}{3} = \frac{2}{3}$.

Validity of the continuous random variable

Definition (Continuous Random Variable)

X is a continuous random variable if it takes uncountably many values. The density function is denoted

$$f_X(x)$$

Properties of the continuous random variable

A continuous random variable must satisfy

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Example

Can $f_X(x) = 2e^{-x}$ for $x \geq 0$ be a continuous random variable?

No, because $\int_{-\infty}^{\infty} f_X(x) = \int_0^{\infty} 2e^{-x}\, dx = 2$.

Remark

If X is a continuous random variable, then $\mathbb{P}(X = x) = 0$ for any x. We *must* consider the probability that it lies in some interval.

If X is a continuous random variable, it's always defined on some interval (can be $\mathbb{R}$). As a convention, wherever it's not defined we just assume that the density is 0.

Cumulative Distribution Function

Definition (Cumulative Distribution Function)

The CDF $F_X(x)$ is the function given by $\boxed{F_X(x) = \mathbb{P}(X \leq x)}$

Properties of the CDF

The CDF must satisfy the following properties

- $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$
- $F_X(x)$ is non-decreasing
- Right-continuous

Important property of the CDF

Assuming $a < b$,

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$$

Where people lose marks

The CDF isn't just defined over some small interval. It's defined over all of $\mathbb{R}$.

Remark

That was not necessarily the most efficient way of doing the problem.

We could've recycled some earlier computations along the way.

CDF of a continuous random variable

Lemma

$$\frac{d}{dx}F_X(x) = f_X(x)$$

Quantiles

Definition (Quantiles)

The k-th quantile of X is the solution to the equation

$$F_X(x) = k.$$

Example: The median is just the value of x such that $F_X(x) = \frac{1}{2}$.

Useful remark

The function Q_X is just the inverse function of F_X.

Example

Find the lower quartile (25% quantile) of the $\text{Exp}\left(\frac{1}{2}\right)$ distribution.

Quantiles

Example

Find the lower quartile (25% quantile) of the $\text{Exp}\left(\frac{1}{2}\right)$ distribution.

The density function is $f_X(x) = \frac{1}{2}e^{-x/2}$ for $x \geq 0$. We're only interested in the CDF for $x \geq 0$.

$$F_X(x) = \int_0^x \frac{1}{2}e^{-t/2}\,dt = 1 - e^{-x/2}$$

(for $x \geq 0$).

Quantiles

Example

Find the lower quartile (25% quantile) of the $\text{Exp}\left(\frac{1}{2}\right)$ distribution.

Setting $F_X(x) = \frac{1}{4}$ gives

$$\begin{aligned} \frac{1}{4} &= 1 - e^{-x/2} \\ e^{-x/2} &= \frac{3}{4} \\ \frac{x}{2} &= -\ln \frac{3}{4} \\ x &= 2 \ln \frac{4}{3} \end{aligned}$$

Expectation

Definition (Expected Value)

For a discrete random variable X, its expectation is

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

For a continuous random variable X, its expectation is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

Expectation

Definition (Expected Value after Transform)

For a discrete random variable X.

$$\mathbb{E}[g(X)] = \sum_{\text{all } x} g(x)\,\mathbb{P}(X = x).$$

For a continuous random variable X,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\,f_X(x)\,dx.$$

Properties of the Expectation

Theorem (Properties of taking expectation)

- $\mathbb{E}[aX] = a\mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $\mathbb{E}[1] = 1$

Critical misassumption

In general, for any function f,

$$\mathbb{E}[f(X)] \neq f(\mathbb{E}[X])$$

Variance and Standard Deviation

Let $\mathbb{E}[X] = \mu$

Definition (Variance)

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

Theorem (Variance Formula)

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

Definition (Standard Deviation)

$$\text{SD}(X) = \sigma_X = \sqrt{\text{Var}(X)}$$

Variance and Standard Deviation

Example

Prove the variance formula from the definition

$$\begin{aligned}
\mathbb{E}\left[(X-\mu)^2\right] &= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2\mathbb{E}[1] \\
&= \mathbb{E}[X^2] - 2\mu\mu + \mu^2 \\
&= \mathbb{E}[X^2] - \mu^2
\end{aligned}$$

Properties of the Variance

Theorem (Properties of taking variances)

- $\text{Var}(X + b) = \text{Var}(X)$
- $\text{Var}(aX) = a^2 \text{Var}(X)$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$
- $\text{Var}(1) = 0$

Critical misassumption

In general, for any two random variables X and Y,

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

Expectation Computations

Example

Given the distribution of X below, compute its expectation and standard deviation.

x	0	3	9	27
$\mathbb{P}(X = x)$	0.3	0.1	0.5	0.1

Expectation Computations

Example

Given the distribution of X below, compute its expectation and standard deviation.

x	0	3	9	27
$\mathbb{P}(X = x)$	0.3	0.1	0.5	0.1

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\text{all } x} x\,\mathbb{P}(X = x) \\ &= 0 \times 0.3 + 3 \times 0.1 + 9 \times 0.5 + 27 \times 0.1 \\ &= 7.5\end{aligned}$$

Expectation Computations

Example

Given the distribution of X below, compute its expectation and standard deviation.

x	0	3	9	27
$\mathbb{P}(X = x)$	0.3	0.1	0.5	0.1

$$\mathbb{E}[X] = 7.5$$

$$\begin{aligned}\mathbb{E}[X^2] &= 0^2 \times 0.3 + 3^2 \times 0.1 + 9^2 \times 0.5 + 27^2 \times 0.1 \\ &= 114.3\end{aligned}$$

Expectation Computations

Example

Given the distribution of X below, compute its expectation and standard deviation.

x	0	3	9	27
$\mathbb{P}(X = x)$	0.3	0.1	0.5	0.1

$$\mathbb{E}[X] = 7.5$$
$$\mathbb{E}[X^2] = 114.3$$

$$\sigma_X = \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} = \sqrt{114.3 - 7.5^2} = \sqrt{58.05} \approx 7.619$$

Expectation Inequalities

Theorem (Chebychev's (Second) Inequality)

Let $\mathbb{E}[X] = \mu$ and $SD(X) = \sigma$. Then, *regardless of the distribution* of X,

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

Note that this is an *upper* bound.

Expectation Inequalities

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.

We require some bound for $\mathbb{P}(400 \leq X \leq 600)$. Observe that:

$$\begin{aligned}\mathbb{P}(400 \leq X \leq 600) &= \mathbb{P}(-100 \leq X - 500 \leq 100) \\ &= \mathbb{P}(|X - 500| \leq 100) \\ &= \mathbb{P}(|X - \mu| \leq k\sigma^2)\end{aligned}$$

where $\mu = 500$, $\sigma^2 = 100$ and therefore $\boxed{\sigma = 10}$ and $\boxed{k = 10}$.

Expectation Inequalities

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.

From Chebychev's (second) inequality,

$$\mathbb{P}(|X - \mu| > 10\sigma) < \frac{1}{10^2}$$

$$\therefore 1 - \mathbb{P}(|X - \mu| \leq 10\sigma) < \frac{1}{100}$$

$$\mathbb{P}(400 \leq X \leq 600) > \frac{99}{100}$$

Bernoulli Distributions

Definition

A random variable X follows a $Ber(p)$ distributions if

$$P(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases} \tag{1}$$

Significance of each parameter

p is the probability of success

Usage

Used to model (the likelihood of) something that either does or does not happen

Binomial Distribution

Definition

A random variable X follows a Bin(n,p) distribution if

$$P(X = x) = \binom{n}{p} p^x (1-p)^{n-x}$$

Significance of each parameter

- n is number of trials
- p is probability of success

Usage

Used to model how many successes in total of n Bernoulli trials

Geometric Distribution

Definition

A random variable X follows a Geom(p) distribution if

$$P(X = x) = (1 - p)^(x - 1)p$$

Significance of each parameter

p is the probability of success.

Usage

Used to model how many Bernoull trials we need before we reach the success outcome.

Hypergeometric Distribution

Definition

A random variable X follows a Hyp(N,m,n) distribution if

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \qquad 0 \leq x \leq \min(m, n)$$

Significance of each parameter

- n is the number of times we select the items
- N is the size of the population
- m is the number of items in the populations satisfying some criteria.

Usage

Used to model how likely we choose x out of the m desirable items.

Poisson Distribution

Definition

A random variable X follows a Poisson(λ) distribution if

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

Significance of each parameter

λ is the average number of occurrences of an event

Usage

Used to model events that are rare. Recommended when an occurrence of an event is independent from another occurrence.

Exponential Distribution

Definition

A random variable T follows a exp β distribution if

$$f_t(t) = \frac{1}{\beta} e^{-t/\beta} \qquad t > 0$$

Significance of each parameter

$\beta = \frac{1}{\lambda}$. It is the average time taken until the next occurrence of the event.

Usage

Based off the memory-less property.

Uniform Distribution

Definition

A random variable X follows a $Unif(a, b)$ distribution if

$$f_x(x) = \frac{1}{b-a} \qquad a < x < b$$

Significance of each parameter

a and b are the two endpoints.

Normal Distribution

Definition

A random variable X follows a N(μ, σ^2) distribution if

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

Significance of each parameter

- μ is its mean
- σ^2 is its variance

Definition (Standard Normal Distribution)

If Z$\sim$N(0,1), then Z follows the standard normal distribution.

Transforms

Loose definition

The transformation of a random variable X under some function h is just $h(x)$.

Comparing Distributions - QQ plots

Definition

For two data sets, the plot of their quantiles against each other is called Quantile-Quantile Plot.

Using QQ plots

We seek if the QQ plot between our data and that from a known distribution is linear. If this is the case, then they are linear transformations of each other.

Usage

Given some data, we plot its quantiles against that of $N(0,1)$. If the graph is linear, then the unknown data is also from a normal distribution.

Joint Probability Function

Definition

If X and Y are discrete random variables then the joint probability function of X and Y is

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

the probability that X = x and Y = y.

Bivariate Distribution Tute P3

3. (From a past MATH2801/2901 final exam)

A soccer club has two expensive star players recruited from overseas.

Detailed match records give the bivariate distribution for (X, Y) below, where X is the number of points obtained in the game (2 for a win, 1 for a draw, 0 for a loss), and Y is the number of stars playing in the game.

The table below gives values of $\mathbb{P}(X = x, Y = y)$.

		y		
		0	1	2
x	0	0.15	0.12	0.06
	1	0.09	0.12	0.09
	2	0.06	0.16	0.15

Bivariate Distribution Tute P3

(a) Jono goes to a game hoping to see the club win the game and hoping to see at least one of the stars playing in the game.

What is the chance that what Jono is hoping to see will actually happen?

(b) i. Find the marginal distribution of X, the number of points obtained in a game.

ii. Hence find the expected number of points obtained in a game.

iii. Find the variance of the number of points obtained in a game.

(c) To Jono's delight, when he arrives at the game he finds out that both stars will be playing in the game.

What is the conditional distribution of points obtained in the game, given that both stars are playing?

(d) Are X and Y independent? Why?

(e) What is the covariance between X and Y?

Bivariate Distribution Tute P3 Solution

3. (a) We have

$$\begin{aligned}\mathbb{P}(X=2, Y>0) &= \mathbb{P}(X=2, Y=1)+\mathbb{P}(X=2, Y=2)\\ &= 0.16+0.15\\ &= 0.31.\end{aligned}$$

(b) i. We have

x	0	1	2
$f_X(x)$	0.33	0.30	0.37

ii. We have

$$\mathbb{E}(X) = 0\times 0.33 + 1\times 0.30 + 2\times 0.37 = 1.04.$$

iii. We have $\mathbb{E}(X^2) = 0^2\times 0.33 + 1^2\times 0.30 + 2^2\times 0.37 = 1.78$, hence

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 = 1.38 - 0.84^2 = 0.6984.$$

Bivariate Distribution Tute P3 Solution

(c) We have

x	0	1	2
$f_{X\|Y}(x \mid 2)$	0.2	0.3	0.5

(d) No, because $f_{X|Y}(x \mid 2) \neq f_X(x)$ at $x = 0$ (and at $x = 2$).

(e) We have

$$\begin{aligned}
\mathbb{E}(XY) &= 0 \times 0.48 + 1 \times 0.12 + 2 \times 0.25 + 4 \times 0.15 = 1.22 \\
\mathbb{E}(Y) &= 0 \times 0.3 + 1 \times 0.4 + 2 \times 0.3 = 1. \\
\mathrm{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 1.22 - 1.04 \times 1 = 0.18.
\end{aligned}$$

Joint Density Function

Definition

The joint density function of continuous random variable X and Y a bivariate function $f_{x,y}$ with the property

$$\int \int_A f_{X,Y}(x,y) d_x d_y = P((X,Y) \in A)$$

any (measurable) subset A of $\mathbb{R}^2$.

Joint Cumulative Distribution Function (CDF)

The joint CDF of X and Y is

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$$

$$\begin{cases} \sum_{u \leq x} \sum_{v \leq y} P(X = u, Y = v) & (X \text{ discrete}) \\ \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v) d_u d_v & (X \text{ continuous}) \end{cases}$$

Marginal Probability

Definition

For discrete r.v.s X and Y with mass function P(X = x, Y = y),

$$P(X = x) = \sum_{ally} P(X = x, Y = y)$$

$$P(Y = y) = \sum_{allx} P(X = x, Y = y)$$

Definition

For discrete r.v.s X and Y with mass function P(X = x, Y = y),

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

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

Conditional Functions

Definition

The conditional probability function of X, given Y = y, is

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Definition

The conditional density function of X, given Y=y is,

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

Independence (Alternative method 1)

Lemma (Independence of Random Variable)

Two random variables are independent if and only if

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

i.e can replace density with the CDF

Independence (Alternative method 2)

Lemma (Independence of random variable

Two random variables are independent if and only if

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

or

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

Conditional Expectation and Variance

Definition (Conditional Expectation)

$$E[X|Y=y] = \begin{cases} \sum_{allx} P(X=x|Y=y) & discretecase \\ \int_{-\infty}^{\infty} x f_{X|Y}(x|y) d_x & continuouscase \end{cases} \quad (2)$$

Definition (Conditional Variance)

$$Var(X|Y=y) = E[X^2|Y=y] - (E[X|Y=y)^2$$

(And similarly for Y. Basically, just add the condition to the original formula)

Covariance

Let $E[X] = \mu_x$ and $E[Y] = \mu_y$

Definition (Covariance)

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

Theorem (Covariance Formula)

$$Cov(X, Y) = E[XY] - \mu_x \mu_y$$

Definition (Correlation)

$$Corr(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)} = \frac{Cov(X, Y)}{\sqrt{Var(x)Var(Y)}}$$

Covariance results

Theorem (Further properties of taking variances)

- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Theorem (Properties of taking covariances)

- $Cov(aX + bY, Z) = a^2 Cov(X, Z) + b^2 Cov(Y, Z)$
- $Cov(X, aY + bZ) = a^2 Cov(X, Y) + b^2 Cov(X, Z)$
- $Cov(X, X) = Var(X)$

Theorem (Consequence of zero covariance)

$$Cov(X, Y) = 0 \Leftrightarrow E[XY] = E[X]E[Y]$$

Moment Generating Function

Definition

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

$$m_x(u) = E(e^{uX})$$

for $u \in \mathbb{R}$

Moment Generating Function 2018 Final Q2

a) **[6 marks]** Let $X \sim N(0,1)$ and $Y \sim N(0,1)$ be two independent standard normal random variables and set $W = X + Y$.

i) **[4 marks]** Using the moment generating functions $m_X(u)$ and $m_Y(u)$, determine the moment generating function $m_W(u)$.

ii) **[2 marks]** Hence, or otherwise, deduce that

$$W \sim N(0,2).$$

Moment Generating Function 2018 Final Q2 Solution

part i)

$$\begin{aligned} m_X(u) &= \int_{-\infty}^{\infty} e^{ux} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\ &= e^{\frac{1}{2}u^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-u)^2} dx \\ &= e^{\frac{1}{2}u^2} \end{aligned}$$

Similarly, $m_Y(u) = e^{\frac{1}{2}u^2}$, so $X \sim N(0,1), Y \sim N(0,1)$
If X, Y are independent, $m_{X+Y}(u) = m_X(u) \cdot m_Y(u) = e^{u^2}$
part ii) Using the fact that mgf of normal is in the form $e^{\mu u + \frac{1}{2}\sigma^2 u^2}$

$$\begin{aligned} m_W(u) &= e^{2\cdot(\frac{1}{2}u^2)} \\ W &\sim N(0,2) \end{aligned}$$

Statistics

Definition (Statistic)

For a random sample $X_1, \ldots, X_n$, a statistic is just a function of the sample.

Example (Common Statistics)

- Sample mean: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- Sample median: $X_{0.5}$

The Sample Mean

Theorem (Properties of the Sample Mean)

Let $X_1, \ldots, X_n$ be a random sample, with mean μ and variance σ^2. Then,

$$\mathbb{E}[\overline{X}] = \mu \qquad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

The Sample Mean

Example

Prove that $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$ as stated just now.

$$\begin{aligned}
\text{Var}\left(\overline{X}\right) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\,\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) && \text{(indep.)} \\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

Efficiency of Statistics

Definition (Efficiency)

Let $g(X_1, \ldots, X_n)$ and $h(Y_1, \ldots, Y_m)$ be two distinct **unbiased** statistics.

$g(X_1, \ldots, X_n)$ is more **efficient** than $h(Y_1, \ldots, Y_m)$ if it has smaller variance, i.e.

$$\text{Var}[g(X_1, \ldots, X_n)] < \text{Var}[h(Y_1, \ldots, Y_m)]$$

Remark: This means we can use different statistics, *or* sample differently, to increase efficiency.

Sampling methods

- Simple random sample - Sampling in a so that all possible samples are equally likely. (Can be hard to do in practice)
- Stratified random sample - As above, but dividing into subclasses of samples beforehand (e.g. age)
- Cluster sampling - Sampling in small groups

Experimental Design (2801)

- Observational study - We don't manipulate any variables.
- Experiment - We manipulate some variables and observe what happens to a 'response' variable.

Experimental Design (2801)

Important features to include in experiments:

- Compare - showing a change in one variable influences a change in another (e.g. via placebo)
- Randomise - minimise the influences of other factors (e.g. gender)
- Repetition

Estimators

Let $X_1, ..., X_n$ be a random sample with model $\{f_X(x;\theta) : \theta \in \Theta\}$

Definition (Estimator)

An estimator for the parameter θ, denoted $\hat{\theta}$, is just a real valued function of $X_1, ..., X_n$ of the random sample.

$$\hat{\theta} = g(X_1, ..., X_n)$$

Meaning, fundamentally it's just a statistic.
Basically, we want to narrow our focus to useful estimators.
Because estimators are functions of random variables, the estimator itself is a random variable.

Bias

Remember that θ is a parameter, so it's constant. Whereas $\hat{\theta}$ is an estimator, which is a r.v.

Definition (Bias)

Given an estimator $\hat{\theta}$ for θ, its bias is

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

The estimator is 'unbiased' if $\text{bias}(\hat{\theta}) = 0$.

Significance

An estimator is 'biased' when it has a tendency of estimating *a little bit off* what the actual value of the parameter is. The bias measures how much it tends to be off by.

Bias

Example

Let $X_1, \ldots, X_7$ be a random Poisson(λ) sample, and consider the estimator

$$\hat{\lambda} = \frac{1}{28} \sum_{i=1}^{7} i\, X_i = \frac{X_1 + 2X_2 + \cdots + 7X_7}{28}$$

for λ. Is this estimator unbiased?

We compute:

$$\mathbb{E}[\hat{\lambda}] = \mathbb{E}\left[\frac{X_1 + 2X_2 + \cdots + 7X_7}{28}\right]$$

Bias

Example

Let $X_1, \ldots, X_7$ be a random Poisson(λ) sample, and consider the estimator

$$\hat{\lambda} = \frac{1}{28} \sum_{i=1}^{7} i\, X_i = \frac{X_1 + 2X_2 + \cdots + 7X_7}{28}$$

for λ. Is this estimator unbiased?

We compute:

$$\begin{aligned}
\mathbb{E}[\hat{\lambda}] &= \mathbb{E}\left[\frac{X_1 + 2X_2 + \cdots + 7X_7}{28}\right] \\
&= \frac{1}{28}\mathbb{E}[X_1 + 2X_2 + \cdots + 7X_7] \\
&= \frac{1}{28}\left(\mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \cdots + 7\mathbb{E}[X_7]\right)
\end{aligned}$$

Bias

We compute:

$$\begin{aligned}
\mathbb{E}[\hat{\lambda}] &= \mathbb{E}\left[\frac{X_1 + 2X_2 + \cdots + 7X_7}{28}\right] \\
&= \frac{1}{28}\mathbb{E}[X_1 + 2X_2 + \cdots + 7X_7] \\
&= \frac{1}{28}\left(\mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \cdots + 7\mathbb{E}[X_7]\right) \\
&= \frac{1}{28}\left(\lambda + 2\lambda + \cdots + 7\lambda\right) \\
&= \frac{1}{28} \times 28\lambda = \lambda
\end{aligned}$$

Hence $\text{bias}(\hat{\lambda}) = \lambda - \lambda = 0$ and thus it *is* unbiased.

Standard Error (2801 ver)

Definition (Standard Error)

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

Significance

Basically adapted from the significance of the variance; it measures just how much error the estimator is susceptible to.

Steps:

1. Compute $\text{Var}(\hat{\theta})$ the usual way
2. Square root it
3. For the standard error, replace θ with $\hat{\theta}$.

Standard Error (2801 ver)

Example

For the earlier example $\hat{\lambda} = \frac{X_1+2X_2+\cdots+7X_7}{28}$, find $\text{se}(\hat{\lambda})$.

$$\begin{aligned}
\text{Var}(\hat{\lambda}) &= \text{Var}\left(\frac{X_1 + 2X_2 + \cdots + 7X_7}{28}\right) \\
&= \frac{1}{28^2}\left(\text{Var}(X_1) + 4\,\text{Var}(X_2) + \cdots + 49\,\text{Var}(X_7)\right) \qquad \text{(indep.)} \\
&= \frac{1}{28^2} \times 140\lambda = \frac{5}{28}\lambda.
\end{aligned}$$

Therefore $\text{se}(\hat{\lambda}) = \sqrt{\frac{5\hat{\lambda}}{28}}$.

Mean Squared Error

Definition (Mean Squared Error)

Given an estimator $\hat{\theta}$ for θ, its mean squared error is

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

Theorem (MSE Formula)

$$\text{MSE}(\hat{\theta}) = [\text{bias}(\hat{\theta})]^2 + \text{Var}(\hat{\theta}).$$

Definition (Estimated Mean Square Error) (2801)

$$\widehat{\text{MSE}}(\hat{\theta}) = [\text{bias}(\hat{\theta})]^2 + [\text{se}(\hat{\theta})]^2.$$

Mean Squared Error formula - Proof

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}\left[\Big((\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta)\Big)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{\theta}] - \theta)^2\right] + 2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\right]
\end{aligned}$$

from expanding the perfect square. Note that $\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])] = \text{Var}(\hat{\theta})$ by definition, and

$$\mathbb{E}\left[(\mathbb{E}[\hat{\theta}] - \theta)\right] = \mathbb{E}[\text{bias}(\hat{\theta})^2] = \text{bias}(\hat{\theta})^2.$$

(Q: Why was I allowed to take off the expected value brackets?)

Mean Squared Error formula - Proof

As for the leftover bit:

$$2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\right] = 2\left(\mathbb{E}[\hat{\theta}] - \theta\right)\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right]$$

...but

$$\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right] = \mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] = 0.$$

Make sure to remember all your properties of the expected value!

Mean Squared Error formula - "Proof"

As for the leftover bit:

$$2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\right] = 2\left(\mathbb{E}[\hat{\theta}] - \theta\right)\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right]$$

...but

$$\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right] = \mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] = 0.$$

Make sure to remember all your properties of the expected value!

Mean Squared Error

Example

For the earlier example $\hat{\lambda} = \frac{X_1+2X_2+\cdots+7X_7}{28}$, find $\text{MSE}(\hat{\lambda})$.

$$\text{MSE}(\hat{\lambda}) = \text{Var}(\hat{\lambda}) + \text{bias}(\hat{\lambda})^2 = \frac{5\lambda}{28} + 0^2 = \frac{5\lambda}{28}.$$

"Better" Estimators

Significance of MSE

Demonstrates a *trade-off* between the variance and the bias.

Better estimators in the MSE sense

Between two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, $\hat{\theta}_1$ is better (w.r.t. MSE), at some specific value of θ, if

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

"Better" Estimators

Example

Let $\hat{\lambda}_1$ be the estimator that we found earlier, with $\text{MSE}(\hat{\lambda}_1) = \frac{5\lambda}{28}$. Now let $\hat{\lambda}_2 = \overline{X}$. For what values of λ is λ_2 better than λ_1?

We can compute:

$$\text{bias}(\hat{\lambda}_2) = 0$$

$$\text{Var}(\hat{\lambda}_2) = \frac{\lambda}{7}$$

$$\therefore \text{MSE}(\hat{\lambda}_2) = \frac{\lambda}{7}$$

"Better" Estimators

Example

Let $\hat{\lambda}_1$ be the estimator that we found earlier, with $\text{MSE}(\hat{\lambda}_1) = \frac{5\lambda}{28}$. Now let $\hat{\lambda}_2 = \overline{X}$. For what values of λ is λ_2 better than λ_1?

$$\text{MSE}(\hat{\lambda}_2) = \frac{\lambda}{7}$$

Solving $\text{MSE}(\hat{\lambda}_2) < \text{MSE}(\hat{\lambda}_1)$ gives

$$\frac{\lambda}{7} > \frac{5\lambda}{28} \implies \lambda < 0.$$

Application - Sample Proportion

Theorem (Properties of the Sample Mean)

Let $X_1, \ldots, X_n$ be a random sample from the Ber(p) distribution. Then the sample proportion $\hat{p} = \frac{\text{No. of successes}}{\text{No. of trials}}$ satisfies:

$$\mathbb{E}[\hat{p}] = p$$

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Consistency

A sequence of random variables $X_1, \ldots, X_n$ converges in probability to X, i.e. $X_n \xrightarrow{P} X$, if $\forall \varepsilon > 0$,

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

Definition (Consistent Estimator)

$\hat{\theta}_n$ is a consistent estimator for θ if it converges in probability to θ. i.e.

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

Verifying that an estimator is consistent

Theorem (Sufficient criteria for consistency)

If

$$\lim_{n\to\infty} \text{MSE}(\hat{\theta}_n) = 0$$

then $\hat{\theta}_n$ is a consistent estimator for θ.

Quick example: Consider the mean proportion $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for μ. Then

$$\text{MSE}(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n) + \text{bias}(\hat{\theta}_n)^2 = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}.$$

Clearly $\lim_{n\to\infty} \text{MSE}(\hat{\theta}_n) = 0$ so the sample mean is a consistent estimator for μ.

Equivariance

Definition (Equivariance Estimator)

$\hat{\theta}_n$ is an equivariant estimator for θ if $g(\hat{\theta}_n)$ is an estimator for $g(\theta)$.

(Only really useful for the MLE.)

Asymptotic Normality

A sequence of random variables $X_1, \ldots, X_n$ converges in distribution to X, i.e. $X_n \xrightarrow{\mathcal{D}} X$, if

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

Definition (Asymptotically Normal Estimator)

$\hat{\theta}_n$ is an asymptotically normal estimator for θ if

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

You don't need to know how to prove these, just how to use it... (soon)

Maximum Likelihood Estimator (MLE)

Definition (Maximum Likelihood Estimator)

$\hat{\theta}$ is the MLE of θ that maximises the likelihood function $\mathcal{L}(\theta; x)$.

Theorem (Computation of the MLE)

$\hat{\theta}$ *also* maximises the log-likelihood function $l(\theta)$

Properties of the MLE

- Equivariant: $g(\theta_{MLE})$ is also the MLE of $g(\theta)$
- Asymptotically normal
- Consistent (in this course)
- *Asymptotically optimal

The Fisher Information

Definition

Let $l(\theta)$ be the log-likelihood function of a random sample. The Fisher score is just its defined as:

$$S_n(\theta) = l'(\theta; x).$$

Definition

The Fisher information is defined as

$$I_n = -\mathbb{E}[l''(\theta; x)]$$

where we swap out x_i for X_i.

Theorem (Alternate definition of Fisher Information)

$$I_n = \mathbb{E}[(l'(\theta; x))^2]$$

Estimators Example 2017 Finals Q4

a) **[6 marks]** Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson (λ) distribution.

i) Determine the method of moments estimator for λ.

ii) Determine the maximum likelihood estimator for λ.

iii) Determine the Fisher information $\mathrm{I}_n(\lambda)$.

Estimators Example 2017 Finals Q4 Solution

- a) $E(X_i) = Var(X_i) = \lambda$
  so $\hat{\lambda} = \bar{X}$
- b) $Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$l(\lambda) = \sum_{i=1}^{n}(X_i \log \lambda - \lambda - \log X_i!) \quad = \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X_i!$$

To find maximum, $l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$. Thus $\hat{\lambda} = \bar{X}$

Estimators Example 2017 Finals Q4 Solution

- c)

$$
\begin{aligned}
f(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\
\log f(x|\lambda) &= x \log \lambda - \lambda - \log x! \\
\frac{\partial}{\partial \lambda} \log f(x|\lambda) &= \frac{x}{\lambda} - 1 \\
-\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) &= \frac{x}{\lambda^2} \\
\text{so } I_X(\lambda) &= E_\lambda \left(-\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) \right) \\
&= E_\lambda \left(\frac{X}{\lambda^2} \right) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda} \\
I_{\underset{\sim}{X}(\lambda)} &= n \cdot I_{X_i}(\lambda) \\
&= \frac{n}{\lambda}
\end{aligned}
$$

Convergence Theorems

Central Limit Theorem

For a random sample $X_1, \ldots, X_n$ with mean μ and finite variance σ,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Slutsky's Theorem

Suppose we have two sequences of random variables (or random samples) with:

$$X_n \xrightarrow{\mathcal{D}} X \qquad\qquad Y_n \xrightarrow{P} c$$

where c is a constant. Then,

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + c \qquad\qquad X_n Y_n \xrightarrow{\mathcal{D}} cX$$

Central Limit Theorem 2018 Finals Q3

3. [15 marks]Answer this question in a separate book

a) **[7 marks]**

i) **[2 marks]** State the Central Limit theorem in one of the forms given in lectures.

ii) **[5 marks]** Let $X_1, X_2, \ldots, X_{48}$ be i.i.d Uniform(0,1) random variables, and set

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

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

Express your answer in terms of the Φ function.
In your answer include a sketch of a suitable curve with a shaded area corresponding to this probability.

Central Limit Theorem 2018 Finals Q3 Solutions

First note that $X_i \sim \text{Unif}(0,1), i = 1, ..., n,$ where $n = 48$

$$E(X_i) = \frac{a+b}{2} = \frac{0+1}{2} = \frac{1}{2}, Var(X_i) = \frac{(b-a)^2}{12} = \frac{1}{12}$$

Stating the CLT for part (i),

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - E(\bar{X})}{\frac{\sqrt{Var(\bar{X})}}{\sqrt{n}}} \longrightarrow N(0,1)$$

For part (ii)

$$Pr(\bar{X} > 0.55) = 1 - Pr(\bar{X} \leq 0.55)$$

$$Pr(\bar{X} \leq 0.55) = Pr(\frac{\bar{X} - 1/2}{\frac{1/\sqrt{12}}{\sqrt{48}}} \leq \frac{0.55 - 1/2}{\frac{1/\sqrt{12}}{\sqrt{48}}}) \text{ (By CLT)}$$

$$= Pr(Z \leq 1.2) = 0.885$$

$$\therefore Pr(\bar{X} > 0.55) = 0.115$$

The Delta Method

Theorem (Provided on formula sheet!!)

Let $\hat{\theta}_1, \hat{\theta}_2, \ldots$ be a sequence of estimators (or a sequence of random variables) of θ such that

$$\frac{\hat{\theta}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

Then, for any function g that is differentiable at θ, with $g'(\theta) \neq 0$,

$$\frac{g(\hat{\theta}_n) - g(\theta)}{g'(\theta)\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

The Delta Method

Example

Suppose $\hat{\beta}_1, \hat{\beta}_2, \ldots$ is a sequence of $i.i.d.$ Exp(β) random variables. Find the 'asymptotic distribution' of $\ln \hat{\beta}_n$.

From the CLT and the formula sheet:

$$\frac{\hat{\beta}_n - \beta}{\frac{\beta}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

The Delta Method

Example

Suppose $\hat{\beta}_1, \hat{\beta}_2, \ldots$ is a sequence of *i.i.d.* $\text{Exp}(\beta)$ random variables. Find the 'asymptotic distribution' of $\ln \hat{\beta}_n$.

From the CLT and the formula sheet:

$$\frac{\hat{\beta}_n - \beta}{\frac{\beta}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

We know $\beta \in (0, \infty)$, so $\ln$ is differentiable at β. Also $(\ln \beta)'$, i.e. β^{-1}, never equals 0. So we can use the Delta method:

$$\frac{\ln \hat{\beta}_n - \ln \beta}{\frac{1}{\beta} \cdot \frac{\beta}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The Delta Method

Example

Suppose $\hat{\beta}_1, \hat{\beta}_2, \ldots$ is a sequence of $i.i.d.$ Exp(β) random variables. Find the 'asymptotic distribution' of $\ln \hat{\beta}_n$.

Use the properties of the normal distribution!

$$\sqrt{n}\left(\ln \hat{\beta}_n - \ln \beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

The Delta Method

Example

Suppose $\hat{\beta}_1, \hat{\beta}_2, \ldots$ is a sequence of $i.i.d.$ Exp(β) random variables. Find the 'asymptotic distribution' of $\ln \hat{\beta}_n$.

Use the properties of the normal distribution!

$$\sqrt{n}\left(\ln \hat{\beta}_n - \ln \beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

$$\ln \hat{\beta}_n - \ln \beta \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{n}\right)$$

The Delta Method

Example

Suppose $\hat{\beta}_1, \hat{\beta}_2, \ldots$ is a sequence of $i.i.d.$ Exp(β) random variables. Find the 'asymptotic distribution' of $\ln \hat{\beta}_n$.

Use the properties of the normal distribution!

$$\sqrt{n}\left(\ln \hat{\beta}_n - \ln \beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

$$\ln \hat{\beta}_n - \ln \beta \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{n}\right)$$

$$\ln \hat{\beta}_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(\ln \beta, \frac{1}{n}\right)$$

Confidence Intervals (Generic Definition)

In a confidence interval, we put the parameter in the middle, instead of the random variable.

Definition (Confidence Interval)

For a random sample $X_1, \ldots, X_n$ with parameter θ, if

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

for some statistics (estimators) L and U, then a $100(1-\alpha)\%$ confidence interval for θ is

$$(L, U)$$

Note: α is just a percentage!

Confidence Intervals (Generic Definition)

"Example" (Setting $\alpha = 0.05$)

For a random sample $X_1, \ldots, X_n$ with parameter θ, if

$$\mathbb{P}(L < \theta < U) = 0.95$$

for some estimators L and U, then a 95% confidence interval for θ is

$$(L, U)$$

Approximate CI's via Asymptotic Normality

Notation (z-value)

z_α represents the α-th quantile of $Z \sim \mathcal{N}(0,1)$, i.e it satisfies

$$\mathbb{P}(Z < z_\alpha) = \alpha$$

Corollary (Approximate CI)

For a random sample $X_1, \ldots, X_n$ with parameter θ, if $\hat{\theta}_n$ is a consistent and asymptotically normal estimator of θ, then

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

is a $100(1-\alpha)\%$ confidence interval for θ.

Approximate CI's via Asymptotic Normality

"Example" (Setting $\alpha = 0.05$)

For a random sample $X_1, \ldots, X_n$ with parameter θ, if $\hat{\theta}_n$ is a consistent and asymptotically normal estimator of θ, then

$$\left(\hat{\theta}_n - z_{0.975}\ \mathrm{se}(\hat{\theta}), \hat{\theta}_n + z_{0.975}\ \mathrm{se}(\hat{\theta})\right)$$

is a 95% confidence interval for θ.

Approximate CI's via Asymptotic Normality

Example (Adapted from Tutorial)

Consider a random sample $X_1, \ldots, X_n$ from the Poisson(λ) distribution. Take $\hat{\lambda} = \overline{X}$, i.e. use the sample mean as an estimator. Find a 95% approximate confidence interval for λ.

Method 1: Directly use the formula: The sample mean is always consistent and asymptotically normal. Recall that $\text{Var}(X_i) = \lambda$ and since our estimator is the sample mean,

$$\text{Var}(\hat{\lambda}) = \text{Var}(\overline{X}) = \frac{\lambda}{n}$$

so therefore

$$\text{se}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{n}}$$

Approximate CI's via Asymptotic Normality

Example (Adapted from Tutorial)

Consider a random sample $X_1, \ldots, X_n$ from the Poisson(λ) distribution. Take $\hat{\lambda} = \overline{X}$, i.e. use the sample mean as an estimator. Find a 95% approximate confidence interval for λ.

(In case you forgot...) According to *R*,

$$z_{0.975} = \texttt{qnorm(0.975)} = 1.959964$$

so an approximate confidence interval is

$$\left(\overline{X} - 1.96\sqrt{\frac{\hat{\lambda}}{n}}, \overline{X} + 1.96\sqrt{\frac{\hat{\lambda}}{n}} \right)$$

Approximate CI's via Asymptotic Normality

Example (Adapted from Tutorial)

Consider a random sample $X_1, \ldots, X_n$ from the Poisson(λ) distribution. Take $\hat{\lambda} = \overline{X}$, i.e. use the sample mean as an estimator. Find a 95% approximate confidence interval for λ.

Method 2: Derive it on the day: Again, because the sample mean is consistent and asymptotically normal, noting that $\text{Var}(X_i) = \lambda$:

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Therefore

$$\mathbb{P}\left(z_{0.025} < \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} < z_{0.975} \right) = 0.95$$

Approximate CI's via Asymptotic Normality

Note that $\boxed{z_{0.025} = -z_{0.975}}$. Rearrange to make λ the subject:

$$-z_{0.975} < \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} < z_{0.975}$$

$$-\sqrt{\frac{\hat{\lambda}}{n}} z_{0.975} < \hat{\lambda} - \lambda < \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975}$$

$$-\sqrt{\frac{\hat{\lambda}}{n}} z_{0.975} < \lambda - \hat{\lambda} < \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975}$$

$$\hat{\lambda} - \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975} < \lambda < \hat{\lambda} + \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975}$$

Be very careful going from line 2 to line 3!

Approximate CI's via Asymptotic Normality

Example (Adapted from Tutorial)

Consider a random sample $X_1, \ldots, X_n$ from the Poisson(λ) distribution. Take $\hat{\lambda} = \overline{X}$, i.e. use the sample mean as an estimator. Find a 95% approximate confidence interval for λ.

Therefore we can rewrite:

$$\mathbb{P}\left(\hat{\lambda} - \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975} < \lambda < \hat{\lambda} + \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975}\right) = 0.95$$

so a 95% confidence interval is

$$\left(\hat{\lambda} - \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975}, \hat{\lambda} + \sqrt{\frac{\hat{\lambda}}{n}} z_{0.975}\right)$$

(Then just sub everything in.)

Follow-up question

Example (contd. from Tutorial)

For the confidence interval above, suppose that for a sample size of 30 the observed values are:

8 2 5 5 8 6 7 2 4 8 4 2 8 4 5 3 3 6 8 3 6 5 5 4 6 3 7 5 1 5

Under these observed values, what is the relevant confidence interval?

From the calculator, the mean of this data is $\frac{148}{30}$, so subbing $\overline{X} = \frac{148}{30}$ and $n = 30$ gives

$$\left(148/30 - 1.96 \times \sqrt{\frac{148/30}{30}}, 148/30 + 1.96 \times \sqrt{\frac{148/30}{30}}\right)$$

which is approximately (4.1385, 5.7281)

Behaviour of the approximate CI

The confidence interval becomes smaller when we increase n, i.e. add more samples!

A 99% confidence interval is wider than a 95% confidence interval. Why?

The hypotheses

Definition (Null Hypothesis, Alternate Hypothesis)

In the null hypothesis H_0, we claim that our parameter θ takes a particular value, say θ_0.

In the alternate hypothesis H_1, we claim some kind of different dependencies.

The 2801 alternate hypotheses:

- $H_1 : \theta \neq \theta_0$
- $H_1 : \theta > \theta_0$
- $H_1 : \theta < \theta_0$

p-values

Definition (p-value)

The p value tells you how much evidence there is against the null hypothesis.

The smaller the p-value, the more evidence against the null hypothesis there is.

If there's more evidence against the null hypothesis, we reject it.

Set-up of a Hypothesis Test (mostly 2801)

1. State the null and alternate hypotheses
2. State the test statistic, and its distribution if we assume H_0 is true
3. Find the observed value of the test statistic
4. Compute the corresponding p-value
5. Draw a conclusion

Test Statistics in Exact tests (Normal samples)

Suppose we know what the variance σ^2 is. We test $H_0 : \mu = \mu_\theta$.

The null distribution is $\boxed{Z \sim \mathcal{N}(0,1)}$.

H_1 :	Test statistic	p-value	p-value
$\theta \neq \theta_0$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\mathbb{P}(\lvert Z\rvert > \lvert\text{observed value}\rvert)$	$2\,\mathbb{P}(Z > \lvert\text{obs. value}\rvert)$
$\theta > \theta_0$	As above	$\mathbb{P}(Z > \text{observed value})$	
$\theta < \theta_0$	As above	$\mathbb{P}(Z < \text{observed value})$	

Test Statistics in Exact tests (Normal samples)

Suppose we estimate the variance σ^2 via S^2. We test $H_0 : \mu = \mu_0$.

The null distribution is $\boxed{T \sim t_{n-1}}$.

H_1 :	Test statistic	p-value	p-value
$\theta \neq \theta_0$	$\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\mathbb{P}(\lvert T\rvert > \lvert\text{observed value}\rvert)$	$2\,\mathbb{P}(T > \text{obs. value})$
$\theta > \theta_0$	As above	$\mathbb{P}(T > \text{observed value})$	
$\theta < \theta_0$	As above	$\mathbb{P}(T < \text{observed value})$	

Exact Tests (Example)

Example (Course Pack)

A popular brand of yoghurt claims to contain 120 calories per serving. A consumer watchdog group randomly sampled 14 servings of the yoghurt and obtained the following numbers of calories per serving:

160 200 220 230 120 180 140 130 170 190 80 120 100 170

Use this data to test the claim.

Step 1: State the hypotheses.

$$H_0 : \mu = 120 \text{ v.s. } H_1 : \mu \neq 120$$

Exact Tests (Example)

Example (Course Pack)

160 200 220 230 120 180 140 130 170 190 80 120 100 170

Use this data to test the claim.

Step 2: State the test statistic, and its null distribution.

We will consider

$$T = \frac{\overline{X} - \mu}{S/\sqrt{14}}$$

and under H_0,

$$T = \frac{\overline{X} - 120}{S/\sqrt{14}} \sim t_{13}$$

Exact Tests (Example)

Example (Course Pack)

160 200 220 230 120 180 140 130 170 190 80 120 100 170

Use this data to test the claim.

Step 3: Find the observed value of the statistic:

$$\overline{x} = 157.8571$$
$$s = 44.75206$$

so the observed value is

$$\frac{\overline{x} - 120}{s/\sqrt{14}} = \frac{157.8571 - 120}{44.75206/\sqrt{14}} = 3.165183$$

Exact Tests (Example)

Example (Course Pack)

160 200 220 230 120 180 140 130 170 190 80 120 100 170

Use this data to test the claim.

Steps 4/5: Compute the p-value and arrive at a conclusion.

$$
\begin{aligned}
p\text{-value} &= \mathbb{P}\left(\left|\frac{\overline{X}-120}{S/\sqrt{14}}\right| > 3.165183\right) \\
&= 2\mathbb{P}(T > 3.165183) \\
&= 2 \texttt{ * pt(3.165183, df=13, lower.tail=FALSE)} \\
&= 0.00745
\end{aligned}
$$

Strong evidence against H_0. The company lied to us...

Rejection Region

Definition (α-level)

The α-level, sets a standard upon which we reject H_0.

Definition (Rejection region)

Under an α-level, we reject H_0 if our observed value lies in the relevant rejection region.

Test Statistics in Exact tests (Normal samples)

Suppose we know what the variance σ^2 is. We test $H_0 : \mu = \mu_0$.

The null distribution is $\boxed{Z \sim \mathcal{N}(0,1)}$.

H_1 :	Test statistic	Rejection region
$\theta \neq \theta_0$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\left\{\lvert\text{observed value}\rvert > z_{1-\frac{\alpha}{2}}\right\}$
$\theta > \theta_0$	As above	$\left\{\text{observed value} > z_{1-\frac{\alpha}{2}}\right\}$
$\theta < \theta_0$	As above	$\left\{\text{observed value} < z_{1-\frac{\alpha}{2}}\right\}$

Test Statistics in Exact tests (Normal samples)

Suppose we estimate the variance σ^2 via S^2. We test $H_0 : \mu = \mu_0$.

The null distribution is $\boxed{T \sim t_{n-1}}$.

H_1 :	Test statistic	Rejection region
$\theta \neq \theta_0$	$\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\left\{\lvert\text{observed value}\rvert > t_{n-1,1-\frac{\alpha}{2}}\right\}$
$\theta > \theta_0$	As above	$\left\{\text{observed value} > t_{n-1,1-\frac{\alpha}{2}}\right\}$
$\theta < \theta_0$	As above	$\left\{\text{observed value} < t_{n-1,1-\frac{\alpha}{2}}\right\}$

Hypothesis Testing Example 2016 Finals

b) A random sample of 100 postal employees found that the average time the employees had worked at the postal service was $\overline{x} = 7$ years with a (sample) standard deviation of $s = 2$ years. Does this provide convincing evidence that the mean time of employment (in years) has changed from the value of 7.5 that was true 20 years ago?

Carry out an appropriate hypothesis test to answer this question by answering the following:

i) State the null and alternative hypotheses.

ii) State any assumption(s) you are making in order to carry out this hypothesis test.
State what you could do, if anything, to check the plausibility of your assumption(s).

iii) State the formula for the test statistic and the null distribution.

iv) Compute the observed value of the test statistic.

v) Give an expression for the P-value, and then estimate it as best you can from the tables provided.

vi) Write a relevant conclusion concerning the mean time of employment of postal employees.

Hypothesis Testing Example 2016 Finals Solution

- i) $H_0 : \mu = 7.5, H_1 : \mu \neq 7.5$
- ii) This is approximately t distributed with mean 7 and sample standard deviation 2. We can also use the approximation to treat this as Normal distribution with mean 7 and standard deviation 2. We can check this by checking independence assumptions. However, for the sake of completeness, we will use t distribution for now, but this question is doable by normal distribution as well.
- iii) Consider $T = \frac{\bar{X}-\mu}{s/\sqrt{100}}$
  under H_0, $T = \frac{7-7.5}{2/\sqrt{100}} \sim t_9 9$
- iv) Observed value is $\frac{7-7.5}{2/\sqrt{100}} = -2.5$
- v) p-value is

$$Pr\left(\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| > -2.5\right) = 2Pr(T > -2.5)$$

Using R studio, 2*pt(2.5,00,lower.tail=FALSE), answer is 0.0140626. Therefore reject H_0 the mean time of employment has changed.

Acknowledgements

- Alex Zhu - Collecting Course materials and Logistics
- Rui Tong - Slide Materials
- Jack Stephens, David Be Olmedo, - Collecting Course materials
- Danica Yong - Question Preparation

GOOD LUCK IN YOUR FINAL EXAM!