ST2132

AY23/24 SEM 1

github/jovyntls

01. PROBABILITY

- **probability** of an event → the limiting relative frequency of its occurrence as the experiment is repeated many times
- the **realisation** x is a constant, and X is a generator
  - running r experiments gives us r realisations $x_1, \dots, x_r$

Expectation

discrete:
(mass function)

$$E(X) := \sum_{i=1}^{n} x_i p_i$$

continuous:
(density function)

$$E(X) := \int_{-\infty}^{\infty} x f(x)\, dx$$

expectation of a function $h(X)$

$$E\{h(X)\} = \begin{cases} \sum_{i=1}^{n} h(x_i) p_i & X \text{ is discrete} \\ \int_{-\infty}^{\infty} h(x) f(x)\, dx & X \text{ is continuous} \end{cases}$$

Variance

variance, $\operatorname{var}(X) := E\{(X-\mu)^2\}$

standard deviation, $SD(X) := \sqrt{\operatorname{var}(X)}$

- $\operatorname{var}(X) = E(X^2) - E(X)^2$
- $E(X - \mu) = 0$

Law of Large Numbers

mean and variance of r realisations:

$$\bar{x} := \frac{1}{r}\sum_{i=1}^{r} x_i \qquad v := \frac{1}{r}\sum_{i=1}^{r}(x_i - \bar{x})^2$$

LLN: for a function h, as $r \to \infty$,

$$\frac{1}{r}\sum_{i=1}^{r} h(x_i) \to E\{h(X)\}$$
$$\bar{x} \to E(X), \quad v \to \operatorname{var}(X)$$

Monte Carlo approximation

simulate $x_1, \dots, x_r$ from X. by LLN, as $r \to \infty$, the approximation becomes exact

$$E\{h(X)\} \approx \frac{1}{r}\sum_{i=1}^{r} h(x_i)$$

Joint Distribution

(**discrete**) mass function:

$$P(X = x_i, Y = y_j) = p_{ij}$$

(**continuous**) density function:

$$f : \mathbb{R}^2 \to [0, \infty), \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1$$

(**expectation**) for $h : \mathbb{R}^2 \to \mathbb{R}$,

$$E\{h(X,Y)\} = \begin{cases} \sum_{i=1}^{I}\sum_{j=1}^{J} h(x_i, y_j) p_{ij} & X \text{ is discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f(x,y)\, dx\, dy & Y \text{ is continuous} \end{cases}$$

Algebra of RV's

let X, Y be RVs and a, b, c be constants

- $Z = aX + bY + c$ is also an RV
  - $z = ax + by + c$ is a realisation of Z
- linearity of expectation: $E(Z) = aE(X) + bE(Y) + c$
- any theorem about a RV is true about a constant

Covariance

let $\mu_X = E(X)$, $\mu_Y = E(Y)$.

covariance, $\operatorname{cov}(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\}$

- $\operatorname{cov}(X,Y) = E(XY) - \mu_X\mu_Y$
- $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- $\operatorname{cov}(X,X) = \operatorname{var}(X)$
- $\operatorname{cov}(W, aX + bY + c) = a\operatorname{cov}(W,X) + b\operatorname{cov}(W,Y)$
- $\operatorname{var}(aX + bY + c) =$ $a^2\operatorname{var}(X) + b^2\operatorname{var}(Y) + 2ab\operatorname{cov}(X,Y)$
- $\operatorname{var}(\sum_{i=1}^{N} a_i X_i) =$ $\sum_{i=1}^{N} a_i^2 \operatorname{var}(X_i) + 2\sum_{1\le i<j\le N} a_i a_j \operatorname{cov}(X_i, X_j)$

joint $=$ marginal $\times$ conditional distributions

$$\begin{aligned} f(x,y) &= f_X(x) f_Y(y|x) \\ &= f_Y(y) f_X(x|y), \quad x, y \in \mathbb{R} \end{aligned}$$

- $f(x,y)$ is the *joint density*
- $f_X(x), f_Y(y)$ are the *marginal densities*
- $f_Y(\cdot|x)$ is the **conditional** density of Y given $X = x$
- $f_X(\cdot|y)$ is the **conditional** density of X given $Y = y$
- for discrete case, *density* $\equiv$ *probability*, $x \equiv x_i$, $y \equiv y_j$

Independence

- X, Y are independent $\iff \forall x, y \in \mathbb{R}$,
  1. $f(x,y) = f_X(x) f_Y(y)$
  2. $f_Y(y|x) = f_Y(y)$
  3. $f_X(x|y) = f_Y(x)$
- X, Y are independent $\Rightarrow$
  - $E(XY) = E(X)E(Y)$
  - $\operatorname{cov}(X,Y) = 0$

  (the converse does not hold)

Conditional expectation

discrete case

let $f_Y(\cdot|x_i)$ be the conditional pmf of Y given $X = x_i$.

$$E[Y|x_i] := \sum_{j=1}^{J} y_j f_Y(y_j|x_i)$$

$$\operatorname{var}[Y|x_i] := \sum_{j=1}^{J}(y_j - E[Y|x_i])^2 f_Y(y_j|x_i)$$

$E[Y|x_i]$ is like $E(Y)$, with conditional distribution replacing marginal distribution $f_Y(\cdot)$. likewise, $\operatorname{var}[Y|x_i]$ like $\operatorname{var}(Y)$.

continuous case

$$E[Y|x] := \int_{-\infty}^{\infty} y f_Y(y|x)\, dy$$

$$\begin{aligned} \operatorname{var}[Y|x] &:= \int_{-\infty}^{\infty}(y - E[Y|x])^2 f_Y(y|x)\, dy \\ &= E(Y^2|x) - \{E(Y|x)\}^2 \end{aligned}$$

Distributions

if X is iid with expectation μ, SD σ and $S_n = \sum_{i=1}^{n} X_i$,

- $E(S_n) = n\mu$
- $SD(S_n) = \sqrt{n}\sigma$
- variance of sum = sum of variances
  $\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\operatorname{var}(x_i)$

bernoulli

$X \sim Bernoulli(p)$ ⇒ coin flip with probability p

$$E(X_i) = p \qquad \operatorname{var}(X_i) = p(1-p)$$
$$E(S_n) = np \qquad \operatorname{var}(S_n) = np(1-p)$$

binomial

$$X \sim Bin(n,p) \quad \Rightarrow X_i \overset{i.i.d.}{\sim} Bernoulli(p)$$

$$E(X) = np, \quad \operatorname{var}(X) = np(1-p)$$
$$E(X) = \sum_{k=1}^{n} k\binom{n}{k} p^k (1-p)^{n-k}$$
$$\operatorname{cov}(X, n-X) = -\operatorname{var}(X)$$

multinomial

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

- for k outcomes $E_1, \dots, E_k$, $Pr(E_i) = p_i$. For some $1 \le i \le k$, E_i occurs X_i times in n runs.

$(X_1, \dots, X_k)$ has the **multinomial distribution:**

$$Pr(X_1 = x_1, \dots, X_k = x_k) = \binom{n}{x_1, \dots, x_k}\Pi_{i=1}^{k} p_i^{x_i}$$

- where $\binom{n}{x_1,\dots,x_k} = \frac{n!}{x_1!x_2!\dots x_k!}$
  - combinatorially, # of arrangements of $x_1, \dots, x_k$
  - $\sum_{i=1}^{n} x_i = n, \quad x_i \ge 0$

$$E(X) = \begin{bmatrix} np_1 \\ np_2 \\ \vdots \\ np_k \end{bmatrix}, \quad \operatorname{var}(X_i) = np_i(1-p_i)$$

$\operatorname{var}(X) =$ *covariance matrix* M with

$$m_{ij} = \begin{cases} \operatorname{var}(X_i) & \text{if } i = j \\ \operatorname{cov}(X_i, X_j) & \text{if } i \ne j \end{cases}$$

- $\operatorname{cov}(X_i, X_j) < 0$
- $X_i \sim Bin(n, p_i)$
- $X_i + X_j \sim Bin(n, p_i + p_j)$

02. PROBABILITY (2)

Mean Square Error (MSE)

$$MSE = E\{(Y-c)^2\}$$

- predicting Y:
  $MSE = \operatorname{var}(Y) + \{E(Y) - c\}^2$
  - $\min MSE = \operatorname{var}(Y)$ when $c = E(Y)$
- Y and X are correlated:
  $MSE = \operatorname{var}[Y|x] + \{E[Y|x] - c\}^2$
  $MSE = E[(Y-c)^2|x] = E[\{Y - E(Y)\}^2|x]$
  - $\min MSE = \operatorname{var}(Y|x)$ when $c = E[Y|x]$
  - if $c = E(Y)$ instead of $E(Y|x) \Rightarrow$ the MSE increases by $(E(Y|x) - E(Y))^2$

mean MSE

$$\frac{1}{n}\sum_{i=1}^{n}\operatorname{var}[Y|x_i] \approx E\{\operatorname{var}[Y|X]\}$$

random conditional expectations

let X, Y be r.v.s.

- $E[Y|X]$ is a r.v. which takes value $E[Y|x]$ with probability/density $f_X(x)$
- $\operatorname{var}[Y|X]$ is a r.v. which takes value $\operatorname{var}[Y|x]$ with probability/density $f_X(x)$

$$E(E[X_2|X_1]) = E(X_2)$$
$$\operatorname{var}(E[X_2|X_1]) + E(\operatorname{var}[X_2|X_1]) = \operatorname{var}(X_2)$$

CDF (cumulative distribution function)

for r.v. X, let $F(x) = P(X \le x)$

- domain: $\mathbb{R}$; codomain: $[0,1]$

$$F(x) = \int_{-\infty}^{\infty} f(x)\, dx$$

Standard Normal Distribution

$Z \sim N(0,1)$ has density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{z^2}{2}\}, \quad -\infty < z < \infty$$

$$E(Z) = 0, \quad \operatorname{var}(Z) = 1$$

CDF, $\Phi(x) = P(Z \le x) = \int_{-\infty}^{x}\phi(z)\, dz$

- $E(Z) = \int_{-\infty}^{\infty} z\phi(z)\, dz = 0$
  - $E(Z^2) = \int_{-\infty}^{\infty} z^2\phi(z)\, dz = 1$
  - $E(Z^{2k+1}) = 0 \quad \forall k \in \mathbb{Z}_{\ge 0}$

general normal distribution

let $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\nu, \tau^2)$

standardisation: $\frac{X-\mu}{\sigma} \sim N(0,1)$

- summations:
  - for constants $a, b \ne 0$,
    $a + bX \sim N(a + b\mu,\ b^2\sigma^2)$
  - $X + Y \sim N(\mu + \nu,\ \sigma^2 + \tau^2 + 2\operatorname{cov}(X,Y))$
    - $\operatorname{cov}(X,Y) = 0, \quad \Rightarrow \quad X \perp Y$
    - $X \perp Y \quad \Rightarrow \quad X + Y \sim N(\mu + \nu,\ \sigma^2 + \tau^2)$
- for $W = a + bX$,
  - density, $f_W(w) = \frac{d}{dw}F_W(w)$
  - CDF, $F_W(w) = P(X \le \frac{w-a}{b}) = \Phi(\frac{w-a}{b})$

Central Limit Theorem

let $X_1, \dots, X_n$ be iid rv's with expectation μ and SD σ, with $S_n = \sum_{i=1}^{n} X_i$

CLT

as $n \to \infty$, the distribution of the standardised $S_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to $N(0,1)$

- $E(S_n) = n\mu$, $\operatorname{var}(S_n) = n\sigma^2$
- for large n, approximately $S_n \sim N(n\mu, n\sigma^2)$

bernoulli

let $X_i \sim Bernoulli(p)$. then $S_n \sim Binom(n,p)$

- for large n, $S_n = N(np, np(1-p))$
- CLT: standardised $\frac{S_n - np}{\sqrt{n}\sqrt{p(1-p)}} \to N(0,1)$ as $n \to \infty$

Distributions

chi-square (χ^2)

let $Z \sim N(0,1)$. $\Rightarrow$ then $Z^2 \sim \chi^2_1$

- Z^2 has χ^2 distribution with 1 degree of freedom
- degrees of freedom = number of RVs in the sum

$$E(Z^2) = 1, \quad E(Z^4) = 3$$
$$\mathrm{var}(Z^2) = E(Z^4) - \{E(Z^2)\}^2 = 2$$

let $V_1, \dots, V_n$ be iid χ^2_1 RVs and $V = \sum_{i=1}^n V_i$. then

$$V \sim \chi^2_n$$
$$E(V) = n \quad \mathrm{var}(V) = 2n$$

gamma

let $\alpha, \lambda > 0$. The $Gamma(\alpha, \lambda)$ density is

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0$$

where $\Gamma(\alpha)$ is a number that makes density integrate to 1

- χ^2_n RV $\sim Gamma(\frac{n}{2}, \frac{1}{2})$
  - χ^2_n is a special case of Gamma!
  - density of χ^2_1 RV $= \frac{1}{\sqrt{2\pi}} v^{-1/2} e^{-v/2}, \quad v > 0$ $= Gamma(\frac{1}{2}, \frac{1}{2})$
- if $X_1 \sim Gamma(\alpha_1, \lambda)$ and $X_2 \sim Gamma(\alpha_2, \lambda)$ are independent, then $X_1 + X_2 \sim Gamma(\alpha_1 + \alpha_2, \lambda)$

t distribution

let $Z \sim N(0,1)$ and $V \sim \chi^2_n$ be independent.

$$\frac{Z}{\sqrt{V/n}} \sim t_n$$

has a t distribution with n degrees of freedom.

- t distribution is symmetric around 0
- $t_n \to Z$ as $n \to \infty$ (because $\frac{V}{n} \to 1$)

F distribution

let $V \sim \chi^2_m$ and $W \sim \chi^2_n$ be independent.

$$\frac{V/m}{W/n} \sim F_{m,n}$$

has an F distribution with (m, n) degrees of freedom.

- even if $m = n$, still two RVs V, W as they are independent
- for $T \sim t_n$, $T^2 = \frac{Z^2}{V/n} \sim F_{1,n}$

IID Random Variables

let $X_1, \dots, X_n$ be iid RVs with mean $\bar{X}$.

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

S is an estimate of σ

let $X_1, \dots, X_n$ be iid $N(\mu, \sigma^2)$. $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
$$E(\bar{X}) = \mu, \quad \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

more distributions:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

- $\bar{X}$ and S^2 are independent

Multivariate Normal Distribution

let $\boldsymbol{\mu}$ be a $k \times 1$ vector and $\boldsymbol{\Sigma}$ be a *positive-definite* symmetric $k \times k$ matrix.

the random vector $\boldsymbol{X} = (X_1, \dots, X_k)'$ has a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its density function is

$$\frac{1}{(2\pi)^{k/2}\sqrt{det\boldsymbol{\Sigma}}} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}{2}\right)$$

- $E(\boldsymbol{X}) = \boldsymbol{\mu}, \quad \mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$
- for any non-zero $k \times 1$ vector $\boldsymbol{a}$,

$$\boldsymbol{a}'\boldsymbol{X} \sim N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$$

  - $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} > 0$ because $\boldsymbol{\Sigma}$ is positive-definite
  - the product $\boldsymbol{a}'\boldsymbol{X}$ is a scalar (same for $\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$)
- two multinomial normal random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, sizes h and k, are independent if $\mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{0}_{h \times k}$
  - $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ has a multivariate normal distribution; the covariance between $\bar{X}$ and $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ is 0, thus they are independent

03. POINT ESTIMATION

for a variable v in population N,

$$\mu = \frac{1}{N} \sum_{i=1}^N v_i \qquad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (v_i - \mu)^2$$

- $\mu, \ \sigma^2$ are **parameters** (unknown constants)
- a **simple random sample** is used to estimate parameters: individuals drawn from the population at random without replacement

binary variable

for variable v with proportion p in the population,

$$\mu = p, \qquad \sigma^2 = p(1-p)$$

single random draw

for variable v (population of size N, mean μ, variance σ^2), let X be the chosen v-value.

$$E(X) = \mu, \qquad \mathrm{var}(X) = \sigma^2$$

draws with replacement

let $X_1, \dots, X_n$ be random draws with replacement from a population of mean μ and variance σ^2.

$$\text{random sample mean, } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$X_1, \dots, X_n$ are iid with $E(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$

$$E(\bar{X}) = \mu, \ \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

let $x_1, \dots, x_n$ be realisations of n random draws with replacement from the population.

$$\text{sample mean, } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- as $n \to \infty$, $\bar{x} \to \mu$ (LLN)
- sample distribution, x_i has the same distribution as X_i and the population distribution

representativeness

- $X_1, \dots, X_n$ is **representative** of the population
  - as n gets larger, $\bar{X}$ gets closer to μ
- $x_1, \dots, x_n$ are *likely* representative of the population

estimating mean

given data $x_1, \dots, x_n$,

- sample mean, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is an **estimate** of μ
- the error in $\bar{x}$ is $\mu - \bar{x}$; it cannot be estimated
- $\bar{x}$ is a realisation of the **estimator** $\bar{X}$
  - this realisation is used to estimate μ

standard error

the size of error in estimate $\bar{x}$ is roughly $SD(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$

the **standard error** (SE) in $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$

- SE is a constant by definition: $SE = SD(\hat{X}) = \frac{\sigma}{\sqrt{n}}$

estimating σ

intuitive estimate of σ^2, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

$$\text{sample variance, } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$
$$E(s^2) = \sigma^2$$

Point estimation of mean

a population (size N) has unknown mean μ, variance σ^2. for random draws (without replacement) $x_1, \dots, x_n$:

$\bar{x}$ is a realisation of $\bar{X}$, with $E(\bar{X}) = \mu$, $\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$

- μ is estimated as $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- error in $\bar{x}$ is measured by the SE: $\frac{\sigma}{\sqrt{n}} = SD(\bar{X})$
- SE is estimated as $\frac{s}{\sqrt{n}}$

$\Rightarrow$ μ is around $\bar{x}$, give or take $\frac{s}{\sqrt{n}}$

unbiased estimation

- since $E(\bar{X}) = \mu$, $\bar{X}$ is an **unbiased** estimator of μ. $\bar{x}$ is an unbiased estimate.
- S^2 is unbiased for σ^2: $E(S^2) = \sigma^2$
- S is *not* unbiased for σ: $E(S) < \sigma$

Simple random sampling (SRS)

n random draws *without replacement* from a population of mean μ and variance σ^2.

- for $i = 1, \dots, n$, $E(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$
- for $i \neq j$, $\mathrm{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$
- if n/N is relatively large,
  - multiply SE by correction factor $\sqrt{\frac{N-n}{N-1}}$
  - standard error = $\frac{N-n}{N-1}$

$$E(\bar{X}) = \mu, \quad \mathrm{var}(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

- if $n << N$, then SRS is like sampling *with replacement* (treat the data as if they come from IID RVs $X_1, \dots, X_n$)

$$E(\bar{X}) = \mu, \quad \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

estimating proportion p

- in a 0-1 population, $\mu = p, \quad \sigma^2 = p(1-p)$
  - p is estimated as $\bar{x}$ (sample proportion of 1's)
- $SE = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = SD(\hat{p})$
  - estimated by replacing p with $\bar{x}$
- unbiased estimator $\hat{p}$
  - $E(\hat{p}) = p, \quad \mathrm{var}(\hat{p}) = \frac{p(1-p)}{n}, \quad SD(\hat{p}) = SE$
- the estimate of σ is $\hat{\sigma}$, not s
- e.g. if a SRS of size 100 has 78 white balls, $p \approx 0.78 \pm \frac{\sqrt{0.78 \times 0.22}}{\sqrt{100}}$

Gauss Model

Let x_i be a realisation of X_i. $X_1, \dots, X_{100}$ are random draws with replacement from an imaginary population with mean w and variance σ^2. w and σ^2 are parameters (unknown constants).

- $E(X_i) = w, \quad \mathrm{var}\, X_i = \sigma^2$ (since X_i is just 1 draw)
- $E(\bar{X}) = w, \quad \mathrm{var}\, \bar{X} = \frac{\sigma^2}{100}$

04. ESTIMATION (SE, bias, MSE)

let $x_1, \dots, x_n$ be from random draws $X_1, \dots, X_n$ with replacement from a population of mean μ and variance σ^2.

sample mean $\bar{x}$ is an *unbiased estimate* of μ

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

SE = $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ tells us roughly how far $\bar{x}$ is from μ

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

MSE and bias

suppose measurements were from a population with mean $w + b$ where b is a constant: $\quad x_i = w + b + \epsilon_i$

- $E(\bar{X}) = w + b$
- $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
  - $SE = \frac{\sigma}{\sqrt{n}}$ measures how far $\bar{x}$ is from $w + b$, not w
- if $b \neq 0$, then $\bar{x}$ is a biased estimate for w

$$MSE = E\{(\bar{X} - w)^2\} = \frac{\sigma^2}{n} + b^2$$
$$MSE = SE^2 + bias^2$$

as $n \to \infty$, $MSE \to b^2$

conclusion

let θ be a parameter (constant) and $\hat{\theta}$ be an estimator (RV).

$SE = SD(\hat{\theta})$, bias $= E(\hat{\theta}) - \theta$,

MSE = $E\{(\hat{\theta} - \theta)^2 = SE^2 + bias^2\}$

05. INTERVAL ESTIMATION

let $x_1, \dots, x_n$ be realisations of IID RVs $X_1, \dots, X_n$ with unknown $\mu = E(X_i)$ and $\sigma^2 = \mathrm{var}(X_i)$.

sample mean, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

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

standard error, $SE = \frac{s}{\sqrt{n}}$

point estimation: $\mu \approx \bar{x}$, give or take $\frac{s}{\sqrt{n}}$

interval estimation: interval contains μ with some confidence level

interval estimation works well if

- X_i has a normal distribution, for any $n > 1$
- X_i has any other distribution but n is large

normal "upper-tail quantile" z_p

let $Z \sim N(0,1)$. for $0 < p < 1$, let z_p be such that

$$p = \Pr(Z > z_p)$$

- e.g. $z_{0.5} = 0$
- $z_p = (1-p)$-quantile of Z
- for $0 < p < 0.5$, $\Pr(-z_p \leq Z \leq z_p) = 1 - 2p$

(case 1) normal distribution with known σ^2

assume $X_1,\ldots,X_n$ are IID $\sim N(0,1)$ with known σ^2.
for $0<\alpha<1$, $\Pr(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$

confidence interval for μ: the random interval

$$\left(\bar X - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n}, \bar X + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n}\right)$$

contains μ with probability $1-\alpha$,
and produces the realisation $\left(\bar x - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n}, \bar x + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n}\right)$

- $1-\alpha$ is the **confidence level**
- *Proof.* since $\frac{\bar X-\mu}{\sigma/\sqrt n}\sim N(0,1)$,
  - $\Pr(-z_{\frac{\alpha}{2}} \le \frac{\bar X-\mu}{\sigma/\sqrt n} \le z_{\frac{\alpha}{2}}) = 1-\alpha$
  - $\Pr(\bar X - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n} \le \mu \le \bar X + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n}) = 1-\alpha$

(case 2) normal distribution with unknown σ^2

assume $X_1,\ldots,X_n$ are IID $\sim N(\mu,\sigma^2)$ with unknown σ^2.
replace σ with S:

for $0<p<1$, let $t_{p,n}$ be such that
$\Pr(t_n > t_{p,n}) = p$

- $t_{p,n}$ is the *upper* p quartile of the t distribution with n degrees of freedom
  - e.g. $t_{0.1,5}=1.48$ (using $qt(0.9,5)$)
- as $n\to\infty$, $t_{n,p}\to z_p$
- $\frac{\bar X-\mu}{S/\sqrt n}\sim t_{n-1}$
- $\Pr(\bar X - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt n} \le \mu \le \bar X + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt n})$

  the random interval
  $$\left(\bar X - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt n}, \bar X + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt n}\right)$$
  contains μ with probability $1-\alpha$.
- data $x_1,\ldots,x_n$ give realisations $\bar x$ of $\bar X$ and s of S, thus the random interval gives a $(1-\alpha)$-CI for μ:
  $\left(\bar x - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt n}, \bar x + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt n}\right)$

(case 3) general distribution with unknown σ^2

IID $X_1,\ldots,X_n$ with $E(X_i)=\mu$, $\mathrm{var}(X_i)=\sigma^2$ unknown

- for large n, approximately $\frac{S_n-n\mu}{\sqrt n\sigma}\sim N(0,1)$
- since $\frac{S_n-n\mu}{\sqrt n\sigma}\sim N(0,1)=\frac{\bar X-\mu}{\sigma/\sqrt n}$,
  - $\Pr(-z_{\frac{\alpha}{2}} \le \frac{\bar X-\mu}{\sigma/\sqrt n} \le z_{\frac{\alpha}{2}}) \approx 1-\alpha$
  - $\Pr(\bar X - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n} \le \mu \le \bar X + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt n}) \approx 1-\alpha$

  for large n, the random interval
  $$\left(\bar X - z_{\frac{\alpha}{2}}\frac{S}{\sqrt n}, \bar X + z_{\frac{\alpha}{2}}\frac{S}{\sqrt n}\right)$$
  contains μ with probability $\approx 1-\alpha$
- data $x_1,\ldots,x_n$ give realisations $\bar x$ of $\bar X$ and s of S.
- $\left(\bar x - z_{\frac{\alpha}{2}}SE, \bar x + z_{\frac{\alpha}{2}}SE\right)$
  is an *approximate* $(1-\alpha)$-CI for μ.
  - $SE=\frac{s}{\sqrt n}$
  - for SRS, multiply SE by correction factor $\sqrt{\frac{N-n}{N-1}}$
- contains μ with probability $<1-\alpha$
- probability $\to 1-\alpha$ as $n\to\infty$
- **exception**: for Bernoulli, $\sigma=\sqrt{p(1-p)}$ is not estimated by s, but by replacing p with the sample proportion

06. METHOD OF MOMENTS

modified notation of mass/density functions:

- **bernoulli**: $f(x|p)=p^x(1-p)^{1-x}, \quad x=0,1$
  - parameter space is $(0,1)$
- **poisson**: $f(x|\lambda)=\frac{\lambda^x e^{-\lambda}}{x!}, \quad x=0,1,\ldots$
  - parameter space is $\mathbb{R}_+$

parameter estimation

assuming data $x_1,\ldots,x_n$ are realisations of IID RVs $X_1,\ldots,X_n$ with mass/density function $f(x|\theta)$, where θ is unknown in parameter space Θ.

- 2 methods to estimate θ :
  - method of moments (MOM)
  - method of maximum likelihood (MLE)
- for both:
  - the estimate of θ is a realisation of an estimator $\hat\theta$
  - SE is $SD(\hat\theta)$
  - bias is $E(\hat\theta)-\theta$
- parameter space Θ: set of values that can be used to estimate the real parameter value θ

Moments of an RV

the k-th moment of an RV X is
$\mu_k=E(X^k), \quad k=1,2,\ldots$

estimating moments

let $X_1,\ldots,X_n$ be IID with the same distribution as X.

the k-th sample moment is
$\hat\mu_k=\frac1n\sum_{i=1}^n X_i^k$

- $E(\hat\mu_k)=\mu_k$ ⇒ unbiased estimator!
- $\hat\mu_k$ is an estimator of μ_k. For realisations $x_1,\ldots,x_n$, the realisation $\frac1n\sum_{i=1}^n x_i^k$ is an *unbiased* estimate of μ_k.
- hat ($\hat{}$) means estimator (random variable)
  - note that this violates the uppercase=RV, lowercase=(fixed)realisation notation
  - $\bar x=\frac1n\sum_{i=1}^n x_i$

MOM: Poisson

assume $x_1,\ldots,x_n$ are realisations of IID $Poisson(\lambda)$ RVs $X_1,\ldots,X_n$. Let λ be the mean number of emissions per 10 seconds (λ is a parameter).

- let $X\sim Poisson(\lambda)$. $\mu_1=\lambda$. Estimate λ by estimating μ_1 using sample mean $\bar x$, which is an estimator of $\bar X$.
- the MOM estimator is $\hat\lambda=\hat\mu_1=\bar X$
  - the random sample mean
- $\mathrm{var}(X)=\lambda$, $\mathrm{var}(\bar X)=\frac{\lambda}{n}$, SE = SD of estimator $=\sqrt{\frac{\lambda}{n}}$

$$\lambda\approx\bar x\pm\sqrt{\frac{\lambda}{n}}$$

MOM: Bernoulli

Assume $X_1,\ldots,X_n$ are iid $Bernoulli(p)$ RVs.
Finding MOM estimator of p:

- let $X\sim Bernoulli(p)$. ⇒ $\mu_1=p$
- MOM estimator, $\hat p=\hat{\mu_1}=\bar X$
  - random sample proportion of 1's
- SE = SD of estimator $=\sqrt{\mathrm{var}(\hat p)}=\sqrt{\frac{p(1-p)}{n}}$

MOM: Normal

let $X_1,\ldots,X_n$ be iid $N(\mu,\sigma^2)$ with parameters μ,σ^2
for $X\sim N(\mu,\sigma^2)$: parameter space, $\Theta=\mathbb{R}\times\mathbb{R}_+$

1. $\mu_1=\mu, \quad \mu_2=\sigma^2+\mu^2$
2. express $\mu=\mu_1$; $\sigma^2=\mu_2-\mu_1^2$; then add hats
3. MOM estimators:

$$\hat\mu=\bar X \qquad \hat{\sigma^2}=\frac1n\sum_{i=1}^n(X_i-\bar X)^2$$

(to construct CI for σ^2: use S^2 ⇒ since $E(S^2)=\sigma^2$)

MOM: Geometric

let $x_1,\ldots,x_n$ be realisations of IID $Geometric(p)$ RVs $X_1,\ldots,X_n$ with expectation $1/p$.

- for $X\sim Geometric(p)$ ⇒ $E(X)=\frac1p$
  - $\Pr(X=i)=p(1-p)^{i-1}$ for $i=1,2,\ldots$
  - $E(X)=\sum_{i=1}^\infty ip(1-p)^{i-1}=\frac1p$
- $\mu_1=\frac1p$ ⇒ $p=\frac1{\mu_1}$ ⇒ $\hat p=\frac1{\bar X}$
- MOM estimator, $\hat p=\frac1{\bar X}$
  - then MOM estimate $=\frac1{\bar x}$
- SE = $SD(1/\bar X)$ ⇒ use monte carlo to approximate

MOM: Gamma

let $X_1,\ldots,X_n$ be iid $Gamma(\alpha,\lambda)$ RVs with shape parameter $\alpha>0$, rate parameter $\lambda>0$

- $X\sim Gamma(\alpha,\lambda)$, $E(X)=\frac{\alpha}{\lambda}$, $E(X^2)=\frac{\alpha(\alpha+1)}{\lambda^2}$
- express parameters in terms of moments:
  $\mu_1=\frac{\alpha}{\lambda}$, $\mu_2-\mu_1^2=\frac{\alpha}{\lambda^2}$ ⇒ $\lambda=\frac{\mu_1}{\mu_2-\mu_1^2}, \alpha=\lambda\mu_1$
- MOM estimators: $\hat\alpha=\frac{\bar X^2}{\hat\sigma^2}$, $\hat\lambda=\frac{\bar X}{\hat\sigma^2}$

MOM estimators are consistent

let $X_1,\ldots,X_n$ be iid with mass/density $f(x|\theta)$, where $\theta\in\Theta\subset\mathbb{R}$.
Suppose $\theta=g(\mu_1)$ for some *continuous* function g.
Then the MOM estimator is **consistent** (approaches θ with more data)

- the MOM estimator is $\hat\theta=g(\hat\mu_1)$. as $n\to\infty$, $\hat\mu_1\to\mu_1$
- since g is continuous, $\hat\theta\to g(\mu_1)=\theta$
  - **asymptotic unbiasedness**: $E(\hat\theta)\to\theta$

07. MLE

MOM: works through estimating moments - if no formula is available for $SD(\hat\theta)$ or $E(\hat\theta)$, monte carlo can be used
MLE: another estimation method

Likelihood function

let $x_1,\ldots,x_n$ be realisations of iid rvs $X_1,\ldots,X_n$ with density $f(x|\theta)$, $\theta\in\Theta\subset\mathbb{R}^k$.

likelihood function $L:\Theta\to\mathbb{R}_+$ is
$L(\theta)=f(x_1|\theta)\times\cdots\times f(x_n|\theta)=\prod_{i=1}^n f(x_i|\theta)$
loglikelihood function $\ell:\Theta\to\mathbb{R}$ is
$\ell(\theta)=\log L(\theta)=\sum_{i=1}^n\log f(x_i|\theta)$

Maximum Likelihood Estimation (MLE)

$$L(\theta)=\prod_{i=1}^n f(x_i|\theta)$$

- **maximiser** of L → the maximum likelihood estimate of θ
  (a realisation of the MLEstimator $\hat\theta$)
  - maximiser of loglikelihood $\ell=\log L$ over Θ

poisson (log)likelihood/MLE

$Poisson(\lambda): f(x|\lambda)=\frac{\lambda^2e^{-\lambda}}{x!}, \ x=0,1,2,\ldots$

- let $x_1,\ldots,x_n$ be realisations of iid Poisson(λ) RVs $X_1,\ldots,X_n$. the joint probability of data is
  $f(x_1|\lambda)\times\cdots\times f(x_n|\lambda)=\frac{\lambda^{\sum_{i=1}^n x_i}e^{-n\lambda}}{x_1!\ldots x_n!}$
- **likelihood**: probability as a function of only λ
  $L(\lambda)=\frac{\lambda^{\sum_{i=1}^n x_i}e^{-n\lambda}}{x_1!\ldots x_n!}$
  - we can leave out constant factors:
    $L(\lambda)=\lambda^{\sum_{i=1}^n x_i}e^{n\lambda}$
- **loglikelihood:**
  $\ell(\lambda)=(\sum_{i=1}^n x_i)\log\lambda-n\lambda-\sum_{i=1}^n\log(x_i!)$
  - leaving out additive constants:
    $\ell(\lambda)=(\sum_{i=1}^n x_i)\log\lambda-n\lambda$
- **MLE of** $\lambda=\bar x$ (maximiser of $L(\lambda)$)
  - differentiate $\ell(\lambda)$: $\ell'(\lambda)=\frac{\sum_{i=1}^n x_i}{\lambda}-n$
  - $\ell'(\lambda)=0$ ⇒ $\lambda=\bar x$
  - $\ell''(\lambda)<0$ (thus max point)

normal (log)likelihood/MLE

$N(\mu,\sigma^2)$: for $x\in\mathbb{R}$,
$f(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=(2\pi)^{\frac12}\sigma^{-1}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- let $x_1,\ldots,x_n$ be realisations of iid N(μ,σ) RVs $X_1,\ldots,X_n$. the joint probability of data is
  $f(x_1|\lambda)\times\cdots\times f(x_n|\lambda)=\frac{\lambda^{\sum_{i=1}^n x_i}e^{-n\lambda}}{x_1!\ldots x_n!}$
- **likelihood** function: joint density as a function of (μ,σ)

$$L(\mu,\sigma)=f(x_1|\mu,\sigma)\times\cdots\times f(x_n|\mu,\sigma)$$
$$=(2\pi)^{-\frac n2}\sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2}$$

- **loglikelihood:**
  $\ell(\mu,\sigma)=-\frac n2\log(2\pi)-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2$
- **MLE:**
  - MLE of $\mu=\bar x$
  - MLE of $\sigma=\hat\sigma=\sqrt{\frac1n\sum_{i=1}^n(x_i-\bar x)^2}$

Gamma distribution

$Gamma(\alpha,\lambda): \ f(x|\alpha,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \ x>0$

- **log of density:** $\alpha\log\lambda-\log\Gamma(\alpha)+(\alpha-1)\log x-\lambda x$
- **loglikelihood:**

$$n\alpha\log\lambda-n\log\Gamma(\alpha)+(\alpha-1)\sum_{i=1}^n\log x_i-\lambda\sum_{i=1}^n x_i$$

  - if α is known, then $\ell(\lambda)=n\alpha\log\lambda-\lambda\sum_{i=1}^n x_i$
- differentiate ⇒ the ML estimates of (α,λ) satisfy
  $\log(\frac{\alpha}{\bar x})-\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}+\bar y=0, \quad \lambda=\frac{\alpha}{\bar x}$ where
  $\bar y=\frac1n\sum_{i=1}^n\log x_i$
- the **ML estimators** $(\hat\alpha,\hat\lambda)$ satisfy
  $\log(\frac{\alpha}{\bar X})-\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}+\bar Y=0, \quad \lambda=\frac{\alpha}{\bar X}$
  - $\log(\frac{\hat\alpha}{\bar X})-\frac{\Gamma'(\hat\alpha)}{\Gamma(\hat\alpha)}+\bar Y=0, \quad \hat\lambda=\frac{\hat\alpha}{\bar X}$

ML vs MOM

- MOM estimates can always be written in terms of the data (sample moments)
  - ML uses *
- ML has better (smaller) SE and bias than MOM
- ML estimates are functions of $\bar x$ and $\bar y$. MOM never uses $\bar y$

Kullback-Liebler divergence (KL)

let $\mathbf{q} = (q_1, \ldots, q_k)$ and $\mathbf{p} = (p_1, \ldots, p_k)$ be strictly positive probability vectors.

the **KL divergence** between $\mathbf{q}$ and $\mathbf{p}$ is

$$d_{KL}(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^{k} q_i \log(\frac{q_i}{p_i})$$

- $d_{KL}(\mathbf{q}, \mathbf{p}) \geq 0$ (equality $\iff$ $\mathbf{q} = \mathbf{p}$)
- $d_{KL}(\mathbf{q}, \mathbf{p}) \neq d_{KL}(\mathbf{p}, \mathbf{q})$

Multinomial

let $(x_1, \ldots, x_n)$ be strictly positive realisations from $(X_1, \ldots, X_n) \sim Multinomial(n, \mathbf{p})$.

- $L(\mathbf{p}) = \Pr(X_1 = x_1, \ldots, X_k = x_k) = cp_1^{x_1} \ldots p_k^{x_k}$
  $= p_1^{x_1} \ldots p_k^{x_k}$ (simplified)
- $\ell(\mathbf{p}) = x_1 \log p_1 + \cdots + x_k \log p_k$
- maximising ℓ via KL divergence
  - if x is from $X \sim Binom(n, p)$, the MOM and ML estimates are both $\hat{p} = \frac{x}{n}$
    - the MOM estimate of p_i is $q_i = \frac{x_i}{n}$.
  - for any $\mathbf{p}$,
    $\ell(\mathbf{q}) - \ell(\mathbf{p}) = \sum_{i=1}^{k} x_i \log q_i - \sum_{i=1}^{k} x_i \log p_i$
    $= n\, d_{KL}(\mathbf{q}, \mathbf{p}) \geq 0$
    - $\ell(\mathbf{q}) - \ell(\mathbf{p}) = 0 \iff \mathbf{p} = \mathbf{q}$

Hardy-Weinberg equilibrium (HWE)

let θ be the proportion of a.

the population is in **HWE** if

$$f(aa) = \theta^2, \quad f(aA) = 2\theta(1-\theta), \quad f(AA) = (1-\theta)^2$$

- (e.g. genotypes) Under HWE, the number of a alleles in an individual has a $Binom(2, \theta)$ distribution
  - for n randomly chosen people, number of a alleles $(AA, Aa, aa) \sim Multinomial(n, \theta)$

Multinomial ML estimation

for $(X_1, X_2, X_3) \sim Multinomial(n, \mathbf{p})$
where $p_1 = (1-\theta)^2$, $p2 = 2\theta(1-\theta)$, $p_3 = \theta^2$

- $L(\theta) = (1-\theta)^{2x_1} 2^{x_2} \theta^{x_2} (1-\theta)^{x_2} \theta^{2x_3}$
  $= 2^{x_2}\, (1-\theta)^{2x_1+x_2}\, \theta^{x_2+2x_3}$
- $\ell(\theta) = x_2 \log 2 + (2x_1 + x_2) \log(1-\theta) + (x_2 + 2x_3) \log \theta$
- ML estimator: $\hat{\theta} = \frac{X_2 + 2X_3}{2n}$
- SE estimation: $\sqrt{\frac{\theta(1-\theta)}{2n}}$
  - $X_2 + 2X_3$ is the number of a alleles: $Binom(2n, \theta)$
    $\Rightarrow \mathrm{var}(\hat{\theta}) = \frac{\theta(1-\theta)}{2n}$

08. LARGE-SAMPLE DISTRIBUTION OF MLEs

let $X_1, \ldots, X_n$ be iid $Geometric(0.5)$ RVs, with mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
by CLT, $\bar{X}_n$ and $\frac{1}{\bar{X}_n}$ have a normal distribution.

asymptotic normality of ML estimator

let $\hat{\theta}_n$ be the ML estimator of $\theta \in \Theta \subset \mathbb{R}$, based on iid RVs $X_1, \ldots, X_n$ with density $f(x|\theta)$.

for large n, the distribution of $\hat{\theta}_n$ is approximately

$$N(\theta, \frac{\mathcal{I}(\theta)^{-1}}{n})$$

where $\mathcal{I}(\theta)$ is the Fisher information derived from $f(x|\theta)$

- $\hat{\theta}_n$ is asymptotically unbiased (like MOM)
  - $E(\hat{\theta}_n) \neq \theta$ (biased)

Fisher Information

let X have density $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$.

the **Fisher information** is the $p \times p$ matrix

$$\mathcal{I}(\theta) = -E\left[\frac{d^2 \log f(X|\theta)}{d\theta^2}\right]$$

- $\mathcal{I}(\theta)$ is symmetric, with (ij)-entry $-E\left[\frac{\delta^2 \log f(X|\theta)}{\delta\theta_i \delta\theta_j}\right]$
- $\mathcal{I}(\theta)$ measures the information about θ in one sample X.

Asymptotic normality: Bernoulli

$X \sim Bernoulli(p): \; f(x|p) = p^x (1-p)^{1-x}, \; x = 0, 1$

Fisher information

- $\log f(X|p) = X \log p + (1-X) \log(1-p)$
- differentiate $\frac{d}{dp}$: $\frac{X}{p} - \frac{1-X}{1-p}$
- differentiate $\frac{d^2}{dp^2}$: $-\frac{X}{p^2} - \frac{1-X}{(1-p)^2}$
- $\mathcal{I}(p) = -E(\frac{d^2 \log f(X|p)}{dp^2}) = \frac{1}{p(1-p)}$
  - minimised at $p = 0.5$

Asymptotic normality

for $X_1, \ldots, X_n$ iid $Bernoulli(p)$ RVs,
Fisher information in each X_i: $\mathcal{I}(p) = \frac{1}{p(1-p)}$

- ML estimator $\hat{p} = \bar{X}$
- for large n, $\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$
  - $E(\hat{p}) = p, \quad \mathrm{var}(\hat{p}) = \frac{p(1-p)}{n}$

Asymptotic normality: Geometric

$X \sim Geometric(p): \; f(x|p) = p(1-p)^{1-x}$

Fisher information

- $\log f(X|p) = \log p + (X-1) \log(1-p)$
- differentiate $\frac{d}{dp}$: $\frac{1}{p} - \frac{X-1}{1-p}$
- differentiate $\frac{d^2}{dp^2}$: $-\frac{1}{p^2} - \frac{X-1}{(1-p)^2}$
- $\mathcal{I}(p) = -E(\frac{d^2 \log f(X|p)}{dp^2}) = \frac{1}{p(1-p)} + \frac{1}{p^2} = \frac{1}{p^2(1-p)}$

Asymptotic normality

for $X_1, \ldots, X_n$ iid $Geometric(p)$ RVs,
Fisher information in each X_i, $\mathcal{I}(p) = \frac{1}{p^2(1-p)}$

- ML estimator $\hat{p} = \frac{1}{\bar{X}}$
- for large n, $\hat{p} \approx N\left(p, \frac{p^2(1-p)}{n}\right)$
  - $E(\hat{p}) > p$ since $E(\hat{p}) = E(\frac{1}{\bar{X}}) > \frac{1}{E(\bar{X})} = p$
  - likely $\mathrm{var}(\hat{p}) \neq \frac{p^2(1-p)}{n}$

Asymptotic normality: Normal

Fisher information

$X \sim N(\mu, \sigma^2), \theta = (\mu, \sigma)$.

$f(x|p) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \quad x \in \mathbb{R}$

- $\log f(X|p) = \frac{1}{2} \log 2\pi - \log \sigma - \frac{(x-\mu)^2}{2\sigma^2 n}$
  $= c - \log \sigma - \frac{(X-\mu)^2}{2\sigma^2 n}$
- differentiate $\frac{d}{dp}$: $\frac{\delta}{\delta\mu} = \frac{X-\mu}{\sigma^2}$, $\frac{\delta}{\delta\sigma} = -\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}$
- differentiate $\frac{d^2}{dp^2}$: $\begin{bmatrix} \frac{\delta^2}{\delta\mu^2} & \frac{\delta^2}{\delta\mu\delta\sigma} \\ \frac{\delta^2}{\delta\sigma\delta\mu} & \frac{\delta^2}{\delta\sigma^2} \end{bmatrix}$
- $\mathcal{I}(p) = -E(\frac{d^2 \log f(X|\theta)}{d\theta^2}) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$

Asymptotic normality

for $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$ RVs, $\theta = (\mu, \sigma)$,

Fisher information in each $X_i : \mathcal{I}(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$

- ML estimator $\hat{\theta} = \begin{bmatrix} \bar{X} \\ \hat{\sigma} \end{bmatrix}$
- for large n, $\hat{\theta} \approx N\left(\begin{bmatrix} \mu \\ \sigma \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{bmatrix}\right)$

are expectation and variance exact?

- a random variable cannot be exactly normal! (cannot be negative)
  - $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
  - $\hat{\sigma} \sim N(\sigma, \frac{\sigma^2}{2n})$ approximately; $E(\hat{\sigma}) \neq \sigma$

normal data

for $x_1, \ldots, x_n$ IID $N(\mu, \sigma^2)$ RVs with large n,
ML estimates of μ and σ are $\bar{x} = \ldots$ and $\hat{\sigma} = \ldots$

- for approximate variance $\begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{bmatrix}$,
  SEs of $\bar{x}$ and $\hat{\sigma}$ are estimated as $\frac{\hat{\sigma}}{\sqrt{n}}$ and $\frac{\hat{\sigma}}{\sqrt{2n}}$
- approximate $(1-\alpha)$-CI:
  $\mu : \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}\right)$
  $\sigma : \left(\hat{\sigma} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{2n}}, \hat{\sigma} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{2n}}\right)$

Gamma distribution

$X \sim Gamma(\alpha, \lambda)$,
$f(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \; x > 0$
$\log f(X) = \alpha \log \lambda - \log \Gamma(\alpha) + (\alpha - 1) \log X - \lambda X$

let $\psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha)$:
($\psi(\alpha) =$ digamma function, $\psi'(\alpha) =$ trigamma function)

- $\frac{\delta \log f(X)}{\delta\alpha} = \log \lambda - \psi(\alpha) + \log X$
- $\frac{\delta \log f(X)}{\delta\lambda} = \frac{\alpha}{\lambda} - X$
- $\frac{\delta^2 \log f(X)}{\delta\alpha^2} = -\psi'(\alpha)$
- $\frac{\delta^2 \log f(X)}{\delta\lambda^2} = -\frac{\alpha}{\lambda^2}$
- $\frac{\delta^2 \log f(X)}{\delta\alpha\delta\lambda} = \frac{\delta^2 \log f(X)}{\delta\lambda\delta\alpha} = \frac{1}{\lambda}$

$$\mathcal{I}(\alpha, \lambda) = \begin{bmatrix} \psi'(\alpha) & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{bmatrix}$$

Approximate CI with ML estimate

$\hat{\theta}_n$ is the ML estimator of $\theta \in \Theta \subset \mathbb{R}$ based on iid RVs $X_1, \ldots, X_n$. $0 < \alpha < 1$

- for large n, approximately $\hat{\theta}_n \sim N(\theta, \frac{\mathcal{I}(\theta)^{-1}}{n})$.
  for $0 < \alpha < 1$,
  $1 - \alpha \approx \Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{\theta}_n - \theta}{\sqrt{\mathcal{I}(\theta)^{-1}/n}} \leq z_{\frac{\alpha}{2}}\right)$
- the random interval
  $\left(\hat{\theta}_n - z_{\frac{\alpha}{2}} \sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}, \hat{\theta}_n + z_{\frac{\alpha}{2}} \sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}\right)$
  covers θ with probability $\approx 1 - \alpha$
- **MLE**: ML estimate of θ, **SE**: $\sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}$ with θ replaced by MLE
  - approxiate $(1-\alpha) - CI$ for θ is
    $(MLE - z_{\frac{\alpha}{2}} SE, MLE + z_{\frac{\alpha}{2}} SE)$

Scope of asymptotic normality of ML estimators

- for iid normal RVs, let $\hat{\sigma}$ be the ML estimator of σ. then $\hat{\sigma}^2$ is the ML estimator of σ^2
  - both $\hat{\sigma}$ and $\hat{\sigma}^2$ are asymptotically normal
  - $\frac{1}{\hat{\sigma}}$ is also asymptotically normal
- let $\hat{\theta}^n$ be the ML estimator of θ. For strictly increasing or strictly decreasing $h : \Theta \to \mathbb{R}$, $h(\hat{\theta}^n)$ is the ML estimator of $h(\theta)$.
  - for large n, $h(\hat{\theta}^n)$ is approximately normal

population mean vs parameter

for n random draws with replacement from a population with mean μ and variance σ^2,

Estimator	E	var	Distribution
random sample mean, $\hat{\mu}$	μ	$\frac{\sigma^2}{n}$	$\approx$ normal
ML estimator, $\hat{\theta}_n$	$\approx \theta$	$\approx \frac{\mathcal{I}(\theta)^{-1}}{n}$	$\approx$ normal

$\hat{\theta}_n$ is not normal (but may approach normal for large n)

summary

let X have density $f(x|\theta), \theta \in \Theta \subset \mathbb{R}^k$.
The **Fisher information** at θ in X is the $k \times k$ matrix

$$-E\left[\frac{d^2 \log f(X|\theta)}{d\theta^2}\right].$$

let $\hat{\theta}_n$ be the ML estimator of θ based on iid RVs $X_1, \ldots, X_n$ with density $f(x|\theta)$.
For large n, the distribution of $\hat{\theta}_n$ is approximately

$$N\left(\theta, \frac{\mathcal{I}(\theta)^{-1}}{n}\right)$$

$\Rightarrow$ SE can be estimated without monte carlo

$\Rightarrow$ accurate CIs are available

skipped: Fisher information in IID samples; binomial fisher information, MLE; HWE trinomial fisher information

$E(\frac{d \log f(X|\lambda)}{d\lambda}) = 0$

09. HYPOTHESIS TESTING

let $x_1, \ldots, x_n$ be realisations of IID $N(\mu, \sigma^2)$ RVs $X_1, \ldots, X_n$ where μ is a parameter and σ is known.

null hypothesis, $H_0 : \mu = \mu_0$
alternative hypothesis, $H_1 : \mu = \mu_1$

It is believed that $\mu = \mu_0$, but it might be μ_1. 2 methods to test if H_0 should be rejected in favour of H_1 using $\bar{x}$:

- if $\bar{x}$ falls inside the **rejection region**, we reject H_0
  - based on a choice of α (type I error)
- P **value** $\to$ the probability that $\bar{X}$ is more extreme than $\bar{x}$, assuming H_0 is true. (if small, doubt H_0)
  - based on an observed test statistic

if σ is unknown or $x_1, \ldots, x_n \not\sim N(\mu, \sigma^2)$, we can use CLT

Rejection region

$x_1, \ldots, x_n$ are from IID $N(\mu, \sigma^2)$ RVs, with σ known

One-tailed test

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1 > \mu_0$$

under H_0,

$$\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n}), \quad \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\alpha = P_{H_0}(\bar{X} > \mu_0 + c) = \Pr(Z > \frac{c}{\sigma/\sqrt{n}}) \quad \Rightarrow c = z_\alpha$$

- reject H_0 if $\bar{x} - \mu_0 > c$ (for some $c > 0$)
  - $\bar{x}$ is the **test statistic**
  - interval $(\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}},\ \infty)$ is the **rejection region**
    - for a test of **size** α, $c = z_\alpha \frac{\sigma}{\sqrt{n}}$

Hypothesis	$\bar{x} < \mu_0 + c$	$\bar{x} > \mu_0 + c$
$H_0 : \mu = \mu_0$	✓ not reject H_0	$\times(I)$ reject H_0
$H_1 : \mu = \mu_1$	$\times(II)$ not reject H_0	✓ reject H_0

- type I error: rejecting H_0 when it is true
- type II error: not rejecting H_0 when it is false

Size and power

- **size** of a test $\rightarrow$ probability of a Type I error
  - $\alpha := P_{H_0}(\bar{X} > \mu_0 + c)$
  - aka **level**
- **power** of a test $\rightarrow$ $1-$ probability of a Type II error
  - $\beta := P_{H_1}(\bar{X} > \mu_0 + c) \quad \Rightarrow$ power $= 1 - \beta$
  - as $n \rightarrow \infty$, power $\rightarrow 1$
    - increasing power of rejecting H_0
- α and β are both about the same event ($\bar{X}$ is in the rejection region), but calculated under different hypotheses (H_0, H_1)
- $\uparrow c : \ \downarrow \alpha, \downarrow \beta$ ($\downarrow$ type I error, $\uparrow$ type II error)
- commonly $\alpha = 0.05$
  - keep α small since H_0 is the default hypothesis

Two-tailed test

$x_1, \ldots, x_n$ are from iid $N(\mu, \sigma^2)$ RVs, σ known

$$H_0 : \mu = \mu_0, \quad H_0 : \mu = \mu_0, H_1 : \mu = \mu_1 \neq \mu_0$$

- reject H_0 if $|\bar{x} - \mu_0| > c$, for some $c > 0$
  - **rejection region:** $(-\infty,\ \mu_0 - c)$ and $(\mu_0 + c, \infty)$
- $\alpha = P_{H_0}(|\bar{X} - \mu_0| > c) = \Pr\left(|Z| > \frac{c}{\sigma/\sqrt{n}}\right)$
  $= 2\Pr\left(Z > \frac{c}{\sigma/\sqrt{n}}\right)$
- $c = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
  - **rejection region:** $(-\infty,\ \mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) \wedge (\mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \infty)$

Composite hypothesis

- **simple** hypothesis $\rightarrow$ specify a single value ($H_0 : \mu = \mu_0, H_1 : \mu = \mu_1$)
- **composite** hypothesis $\rightarrow$ range of values
  - one-tailed test: $H_0 : \mu = \mu_0,\ H_1 : \mu > \mu_0$
    - rejection region: $(\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}},\ \infty)$
      $\Rightarrow$ no change since it doesn't involve μ_1
  - two-tailed test: $H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$
    - rejection region:
      $(-\infty,\ \mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) \wedge (\mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \infty)$
      $\Rightarrow$ no change since it doesn't involve μ_1
    - if $\bar{x}$ falls *outside* the rejection region, i.e.
      $\mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
      - then H_0 is NOT rejected at level α
      - μ_0 lies in the $(1-\alpha)$-CI for μ
  - as $n \rightarrow \infty$, power $\rightarrow 1$

Hypothesis testing and CI

the $(1-\alpha)$-CI for μ, $\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}\right)$
consists of the values μ_0 for which the test
$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$ is not rejected at level α.

P-value

- P**-value** $\rightarrow$ the probability under H_0 that the random test statistic is more extreme than the observed test statistic
  - small p-value = more ”extreme” (more doubt)
- reject H_0 at level $\alpha \iff P < \alpha$
- generally, P-value for two-tailed test is double that of one-tailed test

formulae for P-value

$H_1 : \mu > \mu_0$

$$P = P_{H_0}(\bar{X} > \bar{x}) = \Pr\left(Z > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

$H_1 : \mu < \mu_0$

$$P = P_{H_0}(\bar{X} < \bar{x}) = \Pr\left(Z < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

$H_1 : \mu \neq \mu_0$

$$P = P_{H_0}(|\bar{X} - \mu_0| > |\bar{x} - \mu_0|) = \Pr\left(|Z| > \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}\right)$$