ST2132
AY23/24 SEM 1
github/jovyntls

worth knowing | given in formula sheet | will be given

01. PROBABILITY

Expectation

for a function h

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

for joint distribution

for $h : \mathbb{R}^2 \to \mathbb{R}$, $E\{h(X,Y)\} =$

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

Variance

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

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

useful cases

- $\mathrm{var}(X-c) = \mathrm{var}(X)$
- $\mathrm{var}(X) = \mathrm{cov}(X,X)$
- $\mathrm{var}(\sum_{i=1}^{N} a_iX_i) =$ $\sum_{i=1}^{N} a_i^2\,\mathrm{var}(X_i) + 2\sum_{1\le i<j\le N} a_ia_j\,\mathrm{cov}(X_i,X_j)$
- variance of sum = sum of variances $\mathrm{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\mathrm{var}(x_i)$

Law of Large Numbers

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

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

$$\bar{x} \to E(X), \quad v \to \mathrm{var}(X)$$

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

Covariance

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

covariance

$$\mathrm{cov}(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\} = E(XY) - \mu_X\mu_Y = \mathrm{cov}(Y,X)$$

$$\mathrm{cov}(W, aX+bY+c) = a\,\mathrm{cov}(W,X) + b\,\mathrm{cov}(W,Y)$$

joint $=$ marginal $\times$ conditional distributions

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

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. $f_X(x|y) = f_X(x)$
- X, Y are independent $\Rightarrow$
  - $E(XY) = E(X)E(Y)$
  - $\mathrm{cov}(X,Y) = 0$

  (the converse does not hold)

Conditional expectation

discrete case

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

continuous case

$$E[Y|x] := \int_{-\infty}^{\infty} yf_Y(y|x)\,dy$$
$$\mathrm{var}[Y|x] := \int_{-\infty}^{\infty}(y-E[Y|x])^2 f_Y(y|x)\,dy = E(Y^2|x) - \{E(Y|x)\}^2$$

Distributions

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

distribution of X	$E(X)$	$\mathrm{var}(X)$
$Bernoulli(p)$	p	$p(1-p)$
$Binomial(n,p)$	np	$np(1-p)$
$Geometric(n,p)$	$1/p$	$(1-p)/p^2$
$Multinomial(n,\mathbf{p})$	$\begin{bmatrix} np_1 \\ np_2 \\ \vdots \\ np_k \end{bmatrix}$	$\mathrm{var}(X_i) = np_i(1-p_i)$; $\mathrm{var}(X) =$ *covariance matrix* M with $m_{ij} = \begin{cases}\mathrm{var}(X_i) & \text{if } i=j \\ \mathrm{cov}(X_i,X_j) & \text{if } i\ne j\end{cases}$

- binomial: n coin flips (bernoulli) with probability p
  - $X\sim Bin(n,p) \quad \Rightarrow X_i \overset{i.i.d.}{\sim} Bernoulli(p)$
  - $P(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$
  - $\mathrm{cov}(X, n-X) = -\mathrm{var}(X)$
- multinomial: tally of k possible outcomes (n events)
  - $\mathrm{cov}(X_i,X_j) < 0$
  - $X_i \sim Bin(n,p_i)$, $X_i + X_j \sim Bin(n, p_i+p_j)$

02. PROBABILITY (2)

Mean Square Error (MSE)

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

$\min MSE = \mathrm{var}(Y)$ when $c = E(Y)$

if Y and X are correlated:

$$MSE = \mathrm{var}[Y|x] + \{E[Y|x]-c\}^2$$

mean MSE

Y is predicted from realisations $x_1,\dots,x_n$

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

random conditional expectations

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

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

CDF (cumulative distribution function)

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

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

$\Rightarrow$ density, $f_W(w) = \frac{d}{dw}F_W(w)$

Standard Normal Distribution

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

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

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

- $E(Z^2) = 1$

general normal distribution

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

Central Limit Theorem

CLT

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

for large n, approximately $S_n \sim N(n\mu, n\sigma^2)$

Distributions

chi-square (χ^2)

let $Z\sim N(0,1)$. $\Rightarrow$ then $Z^2\sim\chi_1^2$ (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 $V_1,\dots,V_n \overset{i.i.d.}{\sim}\chi_1^2$ and $V = \sum_{i=1}^{n} V_i$. then $V\sim\chi_n^2$

$E(V) = n \quad \mathrm{var}(V) = 2n$

gamma

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

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

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

$$E(X) = \frac{\alpha}{\lambda}, \quad \mathrm{var}(X) = \frac{\alpha}{\lambda^2}$$
$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

- 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 $V\sim\chi_n^2$ 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 $V\sim\chi_m^2$ and $W\sim\chi_n^2$ 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

IID Random Variables

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

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

$E(S^2) = \sigma^2$ but $E(S) < \sigma$

more distributions:

$\frac{(n-1)S^2}{\sigma^2}\sim\chi_{n-1}^2$

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

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

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

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 $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$

$E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\quad \mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$

- two multinomial normal random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, sizes h and k, are independent if $\mathrm{cov}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \boldsymbol{0}_{h\times k}$

03. POINT ESTIMATION

for a variable v in population N,

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

- μ, σ^2 are **parameters** (unknown constants)

draws with replacement

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

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

$E(X_i) = \mu$, $\quad \mathrm{var}(X_i) = \sigma^2$

- same distribution: x_i, X_i, population distribution
- the error in $\bar{x}$ is $\mu - \bar{x}$; it cannot be estimated

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

Point estimation of mean

a population (size N) has unknown mean μ, variance σ^2.

standard error

SE is a constant by definition:
$SE = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
point estimation of mean: SE ($\bar{x}$) is estimated as $\frac{s}{\sqrt{n}}$

Simple random sampling (SRS)

n random draws *without replacement* from a population

for $i \neq j$, $\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$

- if n/N is relatively large, account for $\text{cov}(X_i, X_j)$
$$E(\bar{X}) = \mu, \quad \text{var}(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$
- if $n << N$, then SRS is like sampling *with replacement* (treat the data as IID RVs $X_1, \ldots, X_n$)
$$E(\bar{X}) = \mu, \quad \text{var}(\bar{X}) = \frac{\sigma^2}{n}$$

estimating proportion p

- the estimate of σ is $\hat{\sigma}$, not s
- unbiased estimator $\hat{p}$
  - $E(\hat{p}) = p, \quad \text{var}(\hat{p}) = \frac{p(1-p)}{n}, \quad SE = SD(\hat{p})$

04. ESTIMATION (SE, bias, MSE)

for random draws $X_1, \ldots, X_n$ *with replacement*

MSE and bias

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

- $E(\bar{X}) = w + b, \quad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
  - $SE = \frac{\sigma}{\sqrt{n}}$ measures how far $\bar{x}$ is from $w + b$, not w
- if $b \neq 0$, then $\bar{x}$ is a biased estimate for w
- $MSE = E\{(\bar{X} - w)^2\} = \frac{\sigma^2}{n} + b^2$

general case

let θ be a parameter and $\hat{\theta}$ be an estimator (RV).
$SE = SD(\hat{\theta}), \quad$ bias = $E(\hat{\theta}) - \theta$,
$MSE = E\{(\hat{\theta} - \theta)^2\} = SE^2 + bias^2$
as $n \to \infty, \ MSE \to b^2$

05. INTERVAL ESTIMATION

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

point estimation: $\mu \approx \bar{x} \pm \frac{s}{\sqrt{n}}$
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 $Z \sim N(0, 1)$. let z_p be the $(1-p)$-quantile of Z.
$p = \Pr(Z > z_p)$

(case 1) normal distribution with known σ^2

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0, 1)$ with known σ^2.
for $0 < \alpha < 1$, $\Pr(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$

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 (confidence level) $1 - \alpha$

(case 2) normal distribution with unknown σ^2

replace σ with S and use t distribution:
for $0 < p < 1$, let $t_{p,n}$ be such that
$\Pr(t_n > t_{p,n}) = p$
as $n \to \infty, \quad t_{n,p} \to z_p$

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 $1 - \alpha$.

(case 3) general distribution with unknown σ^2

- CLT: for large n, approximately $\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0, 1)$
- since $\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ and $S \approx \sigma$ for large n,

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 $\approx 1 - \alpha$

- 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)
- the estimate of θ is a realisation of an estimator $\hat{\theta}$
- parameter space Θ: set of values that can be used to estimate the real parameter value θ
  - e.g. for $N(\mu, \sigma^2)$, parameter space $\Theta = \mathbb{R} \times \mathbb{R}_+$

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
$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$
$E(\hat{\mu}_k) = E(\frac{1}{n}\sum_{i=1}^n x_i^k) = \mu_k \quad \Rightarrow$ unbiased!

MOM: general

let $X \sim Distribution(\theta)$. to obtain $\bar{x}$ and SE:

1. $\mu = \mu_1, \quad \sigma^2 = \mu_2 - \mu_1^2$
2. express parameters in terms of moments
3. estimate MOM estimator using sample mean $\bar{x}$: $\hat{\theta} = \hat{\mu}_1 = \bar{X}$
4. obtain $SE = SD(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})} = \sqrt{\frac{1}{n}\text{var}(X)}$
$$\theta \approx \bar{x} \pm \sqrt{\frac{\text{var}(X)}{n}}$$

07. MLE

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

loglikelihood function $\ell : \Theta \to \mathbb{R}$ is
$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(x_i|\theta)$$
(can omit additive constants (ℓ)/constant factors (L))

Maximum Likelihood Estimation (MLE)

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

find the value of θ that maximises (log)likelihood:
1. calculate likelihood L, loglikelihood ℓ
2. differentiate loglikelihood ℓ: $\ell'(\theta) = 0$
3. confirm max point: $\ell''(\theta) < 0$

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
- MOM/ML estimates are asymptotically unbiased
  - as $n \to \infty$, $E(\hat{\theta}_n) \to \theta$

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

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

- used to maximise ℓ to find MLE for multinomial
- let $\mathbf{q}$ be the MOM estimate for $\mathbf{p}$. for any $\mathbf{p}$,
$\ell(\mathbf{q}) - \ell(\mathbf{p}) = \sum_{i=1}^k x_i \log q_i - \sum_{i=1}^k x_i \log p_i$
$= n\, d_{KL}(\mathbf{q}, \mathbf{p}) \geq 0$
  - $\ell(\mathbf{q}) - \ell(\mathbf{p}) = 0 \iff \mathbf{p} = \mathbf{q} = \frac{\mathbf{x}}{n}$

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 $p_1 = (1-\theta)^2$, $p_2 = 2\theta(1-\theta)$, $p_3 = \theta^2$
- $L(\theta) = p_1^{x_1} p_2^{x_2} p_3^{x_3} \quad = 2^{x_2}(1-\theta)^{2x_1 + x_2}\theta^{x_2 + 2x_3}$
- $\ell(\theta) = x_2 \log 2 + (2x_1 + x_2)\log(1-\theta) + (x_2 + 2x_3)\log\theta$
- 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: $Binom(2n, \theta)$
  $\Rightarrow \text{var}(\hat{\theta}) = \frac{\theta(1-\theta)}{2n}$

08. LARGE-SAMPLE DISTRIBUTION OF MLEs

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

Fisher Information

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

the **Fisher information** is the $k \times k$ 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.

Approximate CI with ML estimate

$\hat{\theta}_n$ is the ML estimator of θ based on iid RVs $X_1, \dots, X_n$.

- for large n, approximately $\hat{\theta}_n \sim N(\theta, \frac{\mathcal{I}(\theta)^{-1}}{n})$.
- 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 $\approx 1-\alpha$

Scope of asymptotic normality of ML estimators

- 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)

Cramér-Rao inequality

if $\hat{\theta}_n$ is unbiased, then $\mathrm{var}(\hat{\theta}_n) \geq \frac{\mathcal{I}(\theta)^{-1}}{n}$

efficient $\iff$ equality

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

09. HYPOTHESIS TESTING

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

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

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

09.1. Rejection region

one-tailed test: $H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1 > \mu_0$
two-tailed test: $H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1 \neq \mu_0$

1. state hypotheses H_0, H_1.
2. reject H_0 if $\bar{x} - \mu_0 > c$ (or $|\bar{x} - \mu_0| > c$)
3. $c = z_{\alpha(/2)} \frac{\sigma}{\sqrt{n}}$ by normalising $\alpha = P_{H_0}(\bar{X} > \mu_0 + c)$
   - since under H_0, $X \sim N(\mu_0, \frac{\sigma^2}{n})$.
4. **rejection region**: reject H_0 if ...
   - $\bar{x} \in (\mu_0 + c, \infty)$
   - $\bar{x} \in (-\infty, \ \mu_0 - c) \cup (\mu_0 + c, \infty)$

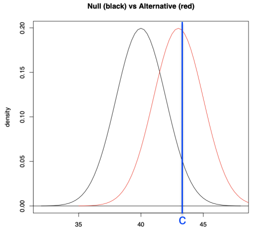
composite H_1: (does not change rejection region)
one-tailed test: $H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0$
two-tailed test: $H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$

Size and power

Hypothesis	$\bar{x} < \mu_0 + c$	$\bar{x} > \mu_0 + c$
H_0	✓ not reject H_0	× (I) reject H_0
H_1	× (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** of a test $\to$ (aka **level**) probability of a Type I error
  - $\alpha := P_{H_0}(\bar{X} > \mu_0 + c)$
  - (for 2-tail) corresponds to a $(1-\alpha)$-CI for μ
- **power** of a test $\to$ $1-$ probability of a Type II error
  - $\beta := P_{H_1}(\bar{X} > \mu_0 + c) \quad \Rightarrow$ power $= 1 - \beta$
  - as $n \to \infty$, power $\to 1$



$\uparrow c : \ \downarrow \alpha, \downarrow \beta$
($\downarrow$ type I err, $\uparrow$ type II err)

09.2. P-value

- **P-value** $\to$ 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

$H_1 : \mu > \mu_0$
$$P = P_{H_0}(\bar{X} > \bar{x}) = \Pr\left(Z > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$
$H_1 : \mu < \mu_0$
$$P = P_{H_0}(\bar{X} < \bar{x}) = \Pr\left(Z < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$
$H_1 : \mu \neq \mu_0$
$$P = P_{H_0}(|\bar{X} - \mu_0| > |\bar{x} - \mu_0|) = \Pr\left(|Z| > \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}\right)$$

10. GOODNESS-OF-FIT

Likelihood Ratio (LR) test

- n iid RVs with density defined by $\theta \in \Omega_1$
- smaller model Ω_0 is **nested** in Ω_1 ($\Omega_0 \subset \Omega_1$)
  - $L_1 \geq L_0 \quad$ (L_0 is the maximum over a subset of L_1)
  - larger $L_1/L_0 \Rightarrow$ poorer fit for smaller model

$H_0 : \theta \in \Omega_0 \qquad H_1 : \theta \in \Omega_1 \backslash \Omega_0$

LR statistic (to test H_0)
$$G = 2\log\left(\frac{L_1}{L_0}\right) = 2(\log L_1 - \log L_0)$$
if $\theta \in \Omega_0$, as $n \to \infty$,
$G \sim \chi^2_{\dim\Omega_1 - \dim\Omega_0}$

LR test: general

1. null hypothesis, H_0 : the tighter model holds
2. LR test statistic,
   $G = 2\log\left(\frac{L_1}{L_0}\right) = 2(\log L_1 - \log L_0)$
3. approximate P-value to χ^2-distribution:
   - $P \approx \Pr\left(\chi^2_{deg} > G\right)$
   - calculate g using *observed count* x_i and *expected count* (under H_0, calculated using ML estimate)
4. high P-value = better fit for tighter model

LR test: Multinomial

let $(X_1, \dots, X_k) \sim Multinomial(n, \mathbf{p})$. then $\mathbf{p} \in \Omega_1$, the set of all positive probability vectors of length k.
let subspace $\Omega_0 = \{(p_1(\theta), \dots, p_k(\theta)) : \theta \in \Theta \subset \mathbb{R}^h\}$ with $\dim\Omega_0 < \dim\Omega_1 = k-1$. to test $H_0 : \mathbf{p} \in \Omega_0$

- $G = 2\sum_{i=1}^k X_i \log\left(\frac{X_i}{np_i(\hat{\theta})}\right)$ (ML estimate of $\mathbf{p}$ is $\frac{\mathbf{x}}{n}$)
  - for Ω_1: $\log L_1 = \sum_{i=1}^k X_i \log(\frac{X_i}{n})$
  - for Ω_0: $\log L_0 = \sum_{i=1}^k X_i \log p_i(\hat{\theta})$
- $P = P_{H_0}(G > g) \approx \Pr(\chi^2_{k-1-\dim\Omega_0} > g)$ for large n.
- to compute g, replace
  - X_i with *observed count* x_i
  - $np_i(\hat{\theta})$ with *expected count* (under H_0) using ML estimate of θ

LR test: Independence

for a population with attributes q and r, let p_{ij} be the population proportion of people with $q = q_i$ and $r = r_j$.
let $(X_{ij} : {}_{1\leq i \leq I, 1 \leq j \leq J}) \sim Multinomial(n, \mathbf{p})$.
H_0 : the two attributes q, r are independent

- $\mathbf{p} \in \Omega_1, \quad \dim\Omega_1 = IJ - 1 = k - 1$.
- if q, r are independent, then $\exists\, q_1, \dots, q_i, r_1, \dots, r_j$ such that $\sum_{i=1}^I q_i = \sum_{j=1}^J r_j = 1$ and $p_{ij} = q_i \times r_j$
- under H_0, for large n, approximately $G \sim \chi^2_{(I-1)(J-1)}$
  - $\dim\Omega_0 = (I-1) + (J-1) = I + J - 2$
  - $\dim\Omega_1 - \dim\Omega_0 = (I-1)(J-1)$
- $G = 2(\log L_1 - \log L_0) = 2\sum_{ij} X_{ij} \log\left(\frac{X_{ij}}{X_{i+}X_{+j}/n}\right)$
  - $\Omega_1 : \log L_1 = \sum_{ij} X_{ij} \log(\frac{X_{ij}}{n})$
  - $\Omega_0 : \log L_0 = \sum_i X_{i+} \log(\frac{X_{i+}}{n}) + \sum_{+j} X_{+j} \log(\frac{X_{+j}}{n})$
- P-value $= \Pr\left(\chi^2_{(I-1)(J-1)} > g\right)$
  - the data x_{ij} are the *observed counts*
  - the data $x_{i+}x_{+j}/n$ are the *expected counts*

LR test: Normal

$X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. to test $H_0 : \mu = 0$:

σ	Ω_1	$\dim\Omega_1$	Ω_0	$\dim\Omega_0$
known	$\mathbb{R}$	1	$\{0\}$	0
unknown	$\mathbb{R} \times \mathbb{R}_+$	2	$\{0\} \times \mathbb{R}_+$	1

under H_0, for large n, approximately $G \sim \chi^2_1$

- **case 1**: σ known
  - $\Omega_0 : \log L_0 = -\frac{n\hat{\mu}^2}{2\sigma^2}, \quad \Omega_1 : \log L_1 = -\frac{n\hat{\sigma}^2}{2\sigma^2}$
  - $G = 2(\log L_1 - \log L_0) = \frac{n\bar{X}^2}{\sigma^2}$
    - if H_0 holds ($\mu = 0$), then $\bar{X} \sim N(0, \frac{\sigma^2}{n})$. for any n, $G \sim \chi^2_1$ exactly.
- **case 2**: σ unknown
  - $\log L_0 = -\frac{n}{2}\log\hat{\mu}_2 - \frac{n}{2}, \ \log L_1 = -\frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}$
  - $G = 2(\log L_1 - \log L_0) = n\log(\frac{\hat{\mu}_2}{\hat{\sigma}^2})$
  - if H_0 holds ($\mu = 0$), for large n, $G \sim \chi^2_1$ approximately