# Introduction to Mathematical Statistics Summary

## Introduction

### Chapter 1 - What this Module is about

① **Why bother:** Stats is distinguishing signal from noise. This module introduces the basic grammar you need! Data ⟶ models [parameters] ⟶ inference [about params]

### Chapter 2 - Basic Concepts

② **Probability Space:** $(\Omega, \mathcal{F}, \mathbb{P})$
- $\Omega$: $\omega \in \Omega$ are the points / elementary events
- $\mathcal{F}$: subsets of $\Omega$ forming a $\sigma$-algebra [take measure theory!]
- $\mathbb{P}$: probability measure, $\mathbb{P}[E]$ assigned to each event $E$

③ **Random variables:** A function $X: \Omega \longrightarrow \mathbb{R}$, measurement, $X(\omega)$ where $\omega$ is the outcome that actually happens. $[X \in B] = \{\omega \in \Omega : X(\omega) \in B\}$, the law is the function $B \longmapsto \mathbb{P}(X \in B)$ - its the distribution.
- Discrete: pmf, $p_X(x) = \mathbb{P}(X = x)$
- Continuous: pdf, $f_X : \mathbb{R} \longrightarrow [0, \infty)$ $\quad \mathbb{P}(X \in B) = \int_B f_X(x)\,dx$
- CDF captures all the info about the distribution, $F_X(x) = \mathbb{P}(X \le x)$

You can calculate expectation, variance, moments $\mathbb{E}[X^n]$ ← nth moment
- $X_1, \dots, X_n$ independent if $\mathbb{P}(X_1 \le a_1, \dots, X_n \le a_n) = \mathbb{P}(X_1 \le a_1) \dots \mathbb{P}(X_n \le a_n)$
- If $A, B$ events, $\mathbb{P}(B) > 0$, then the conditional prob of $A$ given $B$ is

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


"whole set is B, what is size of set A in B?"

④ **Moment Generating Functions:** If $\underbrace{\mathbb{E}[\exp(uX)]}_{= M_X(u)} < \infty \quad \forall u \in (-c, c)$, $c > 0$ MGF
when the MGF exists, it characterises the distribution

**Thm 2.1:** (Uniqueness thm for MGFs). $X, Y$ RVs w/ $M_X(u) = M_Y(u)$ $\forall u \in (-c, c)$ for some $c > 0$. Then $X$ and $Y$ identically distributed

⑤ **Conditional Expectations:** Conditional pmf of $Y$ given $X$ $\quad p_{Y|X=x}(y) = \mathbb{P}(Y = y \mid X = x)$
- If $X \ge 0 \Rightarrow \mathbb{E}[X|Y] \ge 0$ [positivity]
- $\mathbb{E}[aX + bZ \mid Y] = a\,\mathbb{E}[X|Y] + b\,\mathbb{E}[Z|Y]$ [linearity]
- $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ [Tower Property]
- $\mathbb{E}[f(Y)X \mid Y] = f(Y)\,\mathbb{E}[X|Y]$ [taking out what's known]

## Statistical Models

(*) **Calculations to practice** (*)

A statistical model is a parametrized family of distributions $(\mathbb{P}_\theta : \theta \in \Theta)$

### Chapter 3 - A Glossary of Statistical Models

⑥ **Bernoulli(p):** $p \in (0,1)$ $\quad X = \{0, 1\}$ (failure, success)
- pmf: $\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$
- models: wins/losses at roulette, success / fail
- $\mathbb{E}[X] = p$, $\text{var}(X) = p(1-p)$, $M_X(u) = pe^u + 1 - p$ $\quad (\mathbb{E}[e^{uX}] = e^u \cdot p + e^0(1-p))$

⑦ **Binomial(n,p):** $n$ = # trials, $p \in (0,1)$ prob. success
- pmf: $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $\quad k = 0, 1, 2, \dots$

- $\mathbb{E}[X] = np$, $Var(X) = np(1-p)$, $M_X(u) = (1+p(e^u-1))^n$ Pf: $X = \sum_{i=1}^{n} Y_i$ ← Bernoulli treats n independent & use above...
- Models: n iid coin flips

(8) **Geometric (p):** $N \sim Geometric(p)$ is counting time till the first success
- pmf: $\mathbb{P}(N=k) = p(1-p)^{k-1}$ for $k = 1, 2, 3, \ldots$
- $\mathbb{P}(N > n) = (1-p)^n$, $\mathbb{E}[N] = \frac{1}{p}$, $var(X) = \frac{1-p}{p^2}$, $M_X(u) = \frac{pe^u}{1-(1-p)e^u}$
- Memoryless! $\mathbb{P}(N = n+k \mid N > n) = \frac{p(1-p)^{n+k-1}}{(1-p)^n} = p(1-p)^{k-1} = \mathbb{P}(X=k)$ [No effect of previous event]
- Use: $\mathbb{E}[X] = \sum_{k=1}^{\infty} n\,\mathbb{P}(X=n) = \sum_{k=0}^{\infty} \mathbb{P}(X \geq n)$

(9) **NegBinomial(r, p):** n iid bernoulli, rth index $N_r - 1$ occurs.
- pmf: $\mathbb{P}(N=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$ $k = r, r+1, \ldots$
- $\mathbb{E}[N_r] = \frac{r}{p}$, $Var(N_r) = r\frac{(1-p)}{p^2}$, $M_X(u) = \left(\frac{pe^u}{1-(1-p)e^u}\right)^r$

Sum of r geometrics ↳ # sequences length n w/ r ones & ends w/ a 1 [independence + (8)]

Binomial (n,p) w/ $\lambda = np$ & large n

(10) **Poisson ($\lambda$):** counts rare incidents, e.g. # typographical errors in long text
- pmf: $\mathbb{P}(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$
- $\mathbb{E}[X] = \lambda$, $Var(X) = \lambda$, $M_X(u) = e^{\lambda(e^u-1)}$

$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \ldots = \lambda$

$\mathbb{E}[e^{uX}] = \sum_{k=0}^{\infty} e^{uk} \cdot \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^u)^k}{k!} = e^{-\lambda} \cdot e^{\lambda e^u}$

(11) **Categorical (p):** $p \in \mathbb{R}^k$ Models questionnaire responses... $k-1$ dimensional


- pmf: $\mathbb{P}(X=i) = p_i$ $i \in \{1, \ldots, k\}$ $\sum p_i = 1$
- easy to fit to data, but many parameters. Occam's razor!!! Simplest explanation most likely

interesting

$(0, 0, \ldots, 1, \ldots, 0) = e_x$

(12) **Multinomial (n, p):** $p \in \mathbb{R}^k$ lots of categorical variables together. view $X \sim$ categorical (p) as so $X = e_i$ if $X = i$
- pmf: $\mathbb{P}(M = \begin{pmatrix} m_1 \\ \vdots \\ m_k \end{pmatrix}) = \binom{n}{m_1, \ldots, m_k} \prod_{i=1}^{k} p_i^{m_i}$ $(\sum m_i = n)$
- See hard stuff later...

$\binom{n}{m_1, \ldots, m_k} = \frac{n!}{m_1! \ldots m_k!}$

NOTE: measure theory says can't model on all of $\mathbb{R}$

(13) **Uniform (a, b):** model location of random objects on an interval
- $\mathbb{E}[u] = \frac{1}{2}(a+b)$, $var(u) = \frac{1}{12}(b-a)^2$, CDF etc... all easy to calc...
- $\mathbb{E}[e^{pu}] = \frac{1}{b-a}\int_a^b e^{px}\,dx = \frac{e^{pb} - e^{pa}}{p(b-a)}$ can be $\propto$ area / volume in high dim...

(14) **Exponential ($\lambda$):** continuous version of geometric. models radioactive decays
Memoryless / constant failure rate
calculations will need this...
- pdf: $f_T(x) = \lambda e^{-\lambda x}$ for $x > 0$, 0 o/w.
- $\mathbb{E}[T] = \frac{1}{\lambda}$, $Var(T) = \frac{1}{\lambda^2}$
- CDF: $F_T(x) = 1 - e^{-\lambda x}$ $x > 0$
- Memoryless: $\mathbb{P}(T > u+t \mid T > u) = \mathbb{P}(T > t)$
- MGF: $\mathbb{E}[e^{uT}] = \frac{\lambda}{\lambda - u}$ for $u < \lambda$

**Def:** (The Gamma Function)
$$\Gamma(\nu) = \int_0^\infty x^{\nu-1} e^{-x}\,dx, \quad \nu > 0$$
- $\Gamma(\nu+1) = \nu\,\Gamma(\nu)$
- $\Gamma(n) = (n-1)!$

Think substitution to get this form

(15) **Gamma ($\nu, \lambda$):** waiting times when NOT memoryless
- pdf: $f_X(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$, $x > 0$, 0 o/w.
- $\mathbb{E}[T] = \ldots = \frac{\nu}{\lambda}$, $var(T) = \frac{\nu}{\lambda^2}$
- Gamma $(1, \lambda)$ = exponential $(\lambda)$ $\Rightarrow$ $\sum_{i=1}^{n}$ iid exp($\lambda$) is gamma $(n, \lambda)$
- Gamma $(\nu_1, \lambda)$ + Gamma $(\nu_2, \lambda)$ = Gamma $(\nu_1 + \nu_2, \lambda)$
- Mode: compute $\frac{d}{dx} \log f_X(x)$ & assess why max w/ limits

will discuss more distributions when they turn up

## Chapter 4 - Transformations

(16) **Univariate Transformation Theorem:** $g: \mathbb{R} \to \mathbb{R}$ cts'ly diff, strictly increasing, X abs cts R.V. CDF $F_X$, pdf $f_X$. Then $Y = g(X)$ abs cts
$$F_Y(y) = F_X(g^{-1}(y)), \quad f_Y(y) = \frac{1}{g'(g^{-1}(y))} f_X(g^{-1}(y))$$

be careful of range of transformation

Pf: $f_Y(y) = F_Y'(y) = \frac{\partial}{\partial y}(F_X(g^{-1}(y)))$, chain rule & $\frac{\partial}{\partial y}(g(g^{-1}(y))) = 1$ ... ∎

(17) **Standard Normal Distribution:** $Z \sim N(0,1)$, $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ $\forall x \in \mathbb{R}$

recall double integral & polar coordinates trick...

- $\mathbb{E}[Z] = 0$, $\text{var}(Z) = 1$. $F_Z(x) = \Phi(x)$

(18) **General Normal Distribution:** $X \sim N(\mu, \sigma^2)$, $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ $\forall x \in \mathbb{R}$

- Note $X = \mu + \sigma Z$
- CDF $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$
- Linearity of expectation etc: $\mathbb{E}[X] = \mu$, $\text{var}(X) = \sigma^2$
- MGF $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ [complete the square]
- Independence: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, $X \perp Y \Rightarrow X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

  Pf: just use MGF, independence & uniqueness Theorem for MGFs ■

asset prices / exchange rates / amount material

(19) **Log normal Distribution:** To model non-negative data. $Y \sim \text{Log}N(\mu, \sigma^2)$ if $Y = \exp(X)$ w/ $X \sim N(\mu, \sigma^2)$

- use transformation thm to get pdf / cdf.
- MGF does NOT exist! explodes... see notes...
- $\mathbb{E}[Y^n] = \mathbb{E}[e^{nX}] = M_X(n) = \exp(\mu n + \frac{1}{2}\sigma^2 n^2)$

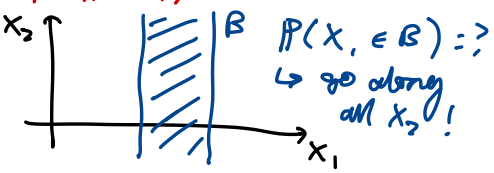
(20) **$\chi^2$ Distribution:** central when assessing statistical procedure. $Y \sim \chi^2_d$ $Y = Z_1^2 + \dots + Z_d^2$ where $Z_i \sim N(0,1)$ iid. $d$ = # degrees of freedom of chi-squared distribution

- If $d=1$, $Y = Z^2$, $F_Y(z) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$,
- $\chi^2_1 = \text{Gamma}(\frac{1}{2}, \frac{1}{2}) \Rightarrow \chi^2_d = \text{Gamma}(\frac{d}{2}, \frac{1}{2})$ ← $d$ independent gamma

Pf: $f_Y(y) = F_Y'(y) = \frac{1}{2\sqrt{y}}\Phi'(\sqrt{y}) + \frac{1}{2\sqrt{y}}\Phi'(-\sqrt{y}) = \frac{1}{\sqrt{y}}\Phi'(\sqrt{y}) = \underbrace{\frac{1}{\sqrt{2\pi}\sqrt{y}}\exp(-\frac{y}{2})}_{\text{Gamma}(\frac{1}{2},\frac{1}{2})}$

(21) **Joint distributions:** when R.V.s interact. $X \in \mathbb{R}^n$, then joint dist is $\mathbb{P}(X \in A)$

- Multivariate distribution function $F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n)$
- $f_X(x) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_X(x_1,\dots,x_n)$
- Multivariate MGF: $M(a) = \mathbb{E}[\exp(a^T X)]$
- To get marginal distribution, just integrate out all else




(22) **Covariance / Covariance matrices:** How much do R.V.s depend on each other?

- $\text{cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Note $X \perp Y \Rightarrow \text{cov}(X,Y) = 0$
- $\text{Var}(aX+bY) = a^2\text{var}(X) + 2ab\,\text{cov}(X,Y) + b^2\text{var}(Y)$

  Pf: $\text{b} = \mathbb{E}[(aX+bY - \mathbb{E}[aX+bY])^2] = \dots$

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

- $-1 \le \text{corr}(X,Y) \le 1$ ← quadratic wrt $a$ $\Rightarrow b^2 - 4ac \le 0$ as Ψ

  Pf: $\text{var}(aX+bY) \ge 0$, $\Rightarrow (2b\,\text{cov}(X,Y))^2 - 4b^2\text{var}(X)\text{var}(Y) \le 0$ ■
- $X_1, \dots, X_n$ independent and finite 2nd moments $\Rightarrow \text{var}(\sum a_i X_i) = \sum a_i^2 \text{var}(X_i)$
- If $X \in \mathbb{R}^n$, $\text{cov}(X)_{ij} = \text{cov}(X_i, X_j)$ [diagonals are variance]

**Example:** Covariance matrix of $X \sim \text{multinomial}(n,p)$ $\in \mathbb{R}^k$

!!! EXAM HINT ??? "Nice example, worth going thru"

$\mathbb{E}[X] = (np_1, \dots, np_k)$, $\text{var}(X) = np_i(1-p_i)$ ∵ $X_i \sim \text{Binomial}(n, p_i)$

$\text{cov}(X_i, X_j)$ ⓐ $i=j \Rightarrow \text{cov} = \text{Var}(X_i) = p_i(1-p_i)$

ⓑ $i \ne j \Rightarrow \text{cov} = \underbrace{\mathbb{E}[X_i X_j]}_{\mathbb{P}(X_i = X_j = 1) = 0} - \mathbb{E}[X_i]\mathbb{E}[X_j] = -p_i p_j$

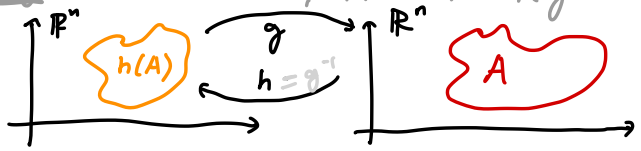
$X$ = sum of $n$ iid categorical($p$)

so

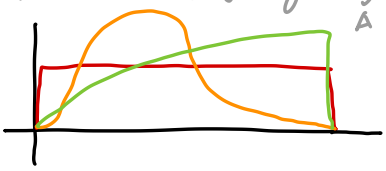
$$\text{cov}(X) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & & \vdots \\ \vdots & & \ddots & \\ -np_kp_1 & \cdots & & np_k(1-p_k) \end{pmatrix}$$

(23) **Multivariate Transformation Theorem:** $X \in \mathbb{R}^n$ abs cts with density $f_X(x)$, $g: \mathbb{R}^n \to \mathbb{R}^n$ is 1:1 & onto & ctsly diff with ctsly diff inverse $h$, then $Y = G(X)$ is abs cts with density $f_Y(y) = J_h(y) f_X(h(y))$, $J_h$ is Jacobian $J_h = |\det(\frac{\partial h}{\partial y})| = \left|\det\begin{pmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{pmatrix}\right|$

Proof: Take $A \subset \mathbb{R}^n$, $P(Y \in A) = P(g(X) \in A) = P(X \in h(A))$




$= \int_{h(A)} f_X(x)\,dx$ (change of variable formula for multiple integrals) $= \int_A f_X(h(y))\, J_h(y)\,dy = \int_A f_Y(y)\,dy$



(24) **Beta$(\alpha, \beta)$:** Very flexible, can model whatever!

- pdf: $f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $x \in (0,1)$, $\alpha, \beta > 0$ (just norm constant)
- If $X \sim \text{Gamma}(\alpha, 1)$, $Y \sim \text{Gamma}(\beta, 1)$, $X \perp Y \Rightarrow \frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$

  Pf: Define $Z = X+Y$, $B = \frac{X}{X+Y}$, $(B, Z) = g(X,Y) = (\frac{X}{X+Y}, X+Y)$, use (23), find marginal ■

(25) **Quick Linear Algebra Review:**

(a) Gram-Schmidt orthogonalization: Given a basis $\{x_1, \dots, x_n\}$ of $\mathbb{R}^n$, orthogonalize by (remove $u_1$ component)
① $u_1 = \frac{x_1}{\|x_1\|_2}$ ② $u_2 = x_2 - (x_2 \cdot u_1)u_1$ ③ normalize $u_2$ ④ repeat.

(b) $V \in \mathbb{R}^{n \times n}$ symmetric $\Rightarrow$ real eigenvalues & orthonormal basis of eigenvectors. $Vu_i = \lambda_i u_i$

(c) If $U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix}$, then $U^T U = I_n$

(d) $V = U \,\text{diag}(\lambda_1, \dots, \lambda_n)\, U^T$

(e) $V$ positive-semidefinite if $x^T V x \geq 0$ $\forall x \in \mathbb{R}^n$ [strict if true $>$]

(f) $V$ has a sqrt $A = U L^{1/2} U^T$, $L = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_n^{1/2})$ [avoids complex $\lambda_i^{1/2}$ w/ $\lambda_i < 0$]

(g) If $V$ symmetric positive semidefinite, then so is $A$.

$A^T A = (U L^{1/2} U^T)^T (U L^{1/2} U^T) = U L^{1/2} \cancel{U^T U} L^{1/2} U^T = U L U^T = V$

(26) **Multivariate Normal:** $X \sim MVN(\mu, V)$ if $\exists \mu \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ s.t. $X = \mu + AZ$, $Z \in \mathbb{R}^m$ with $Z_i \sim N(0,1)$ and $V = AA^T$

- $E[X_i] = E[\mu_i + \sum_j A_{ij} Z_j] = \mu_i$
- $\text{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])] = E[(AZ)_i (AZ)_j] = E[(AZ)(AZ)^T_{ij}]$ $= E[(AZZ^TA^T)_{ij}] = AA^T_{ij} = V_{ij}$
- density: If $V$ has full rank, (use multi-dim mapping thm; $\frac{1}{2\sigma^2}(x-\mu)^2$)

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(V)}} \exp\left(-\tfrac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right)$$

Pf: $X = g(Z) = \mu + AZ \Rightarrow Z = h(x) = A^{-1}(x-\mu) \Rightarrow f_X(x) = J_h(x) f_Z(h(x))$

$\frac{\partial h}{\partial x} = A^{-1} \Rightarrow J_h(x) = |\det(A^{-1})| = \frac{1}{\sqrt{\det(V)}}$ $\because V = AA^T$, know $f_Z(z) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\|z\|_2^2}$

so $f_Z(h(x)) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\|h(x)\|_2^2\right)$

$= \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}(A^{-1}(x-\mu))^T (A^{-1}(x-\mu))\right)$ $\quad (A^{-1})^T A^{-1} = (AA^T)^{-1} = V^{-1}$

$= \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right)$ ■

- MGF: $M_X(u) = E[\exp(u^T X)] = \cdots = \exp(u^T \mu + \frac{1}{2} u^T V u)$

(27) **Fischer's Theorem:** $X_1, \dots, X_n$ independent $N(0, \sigma^2)$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $S_{xx}^2 = \sum_{i=1}^n (x_i - \bar{X}_n)^2$

Then, $\bar{X}_n$ and $S_{xx}^2$ are independent & $\frac{\bar{X}_n}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{1}{\sigma^2} S_{xx}^2 \sim \chi^2_{n-1}$

(wlog take $\sigma = 1$)

Pf: Take $u_1 = (\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})^T$, Gram-Schmidt $\Rightarrow$ orthonormal basis $u_1, \dots, u_n$ of $\mathbb{R}^n$

Set $Y = U^T X$ with $U^T U = I_n$ $Y_i \sim$ iid $N(0,1)$. $Y_1 = \sum \frac{1}{\sqrt{n}} x_i = \frac{\sqrt{n}}{n} \sum x_i = \sqrt{n}\,\bar{X}_n$ (independent of $Y_2, \dots, Y_n$)

Note: $\sum_{i=1}^n Y_i^2 = \|U^T X\|^2 = X^T U U^T X = X^T X = \sum_{i=1}^n x_i^2$

so $S_{xx}^2 = \sum_{i=1}^n (x_i - \bar{X}_n)^2 = \sum_{i=1}^n x_i^2 - n\bar{X}_n = \sum_{i=1}^n Y_i^2 - Y_1 = \sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$ (independent of $Y_1$ $\because$ not $Y_1$ in formula!) ■

**Corollary:** $X_1,\dots,X_n \sim$ iid $N(\mu,\sigma^2)$, Then $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\Rightarrow \bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

Pf: easy! Just work out...

(28) **The F-distribution:** $Y \sim F_{m,n}$ ($m, n$ are degrees of freedom) if $Y = \frac{X_m/m}{X'_n/n}$ where $X_m \sim \chi^2_m$ and $X'_n \sim \chi^2_n$

- **Lemma:** $f_Y(y) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} y^{\frac{m}{2}-1} \left(1+\frac{m}{n}y\right)^{-\frac{1}{2}(n+m)}$

Used in statistical comparison of variance

Pf: Stretched in notes...

---

## Chapter 5 - Approximation Theorems

(29) **Useful inequalities:**

$0 \le a\mathbb{P}(X \le a) \le \mathbb{E}[X]$ ⇒ expectations...

(a) Markov: $X \ge 0$, $\mathbb{P}(X \ge a) \le \frac{\mathbb{E}[X]}{a}$ Pf: $Y = \begin{cases} 1 & \text{if } X \ge a \\ 0 & \text{o/w} \end{cases} \Rightarrow 0 \le aY \le X$

(b) Chebyshev: $X$ R.V. $\mathbb{E}[X]=\mu$, $\text{var}(X)=\sigma^2$, Then $\mathbb{P}(|X-\mu| \ge k) \le \frac{\sigma^2}{k^2}$ Pf: Markov on $|X-\mu|^2$

(30) **Convergence in Random Variables:** $X, X_1, X_2, \dots$ R.V. defined on the same prob space

- Convergence in Quadratic mean (quadratic mean): $X_n \xrightarrow{} X$ in 2-mean if $\mathbb{E}[|X_n - X|^2] \xrightarrow{n\to\infty} 0$
  - Pf: $\mathbb{P}(|X_n-X|>\varepsilon) = \mathbb{P}(|X_n-X|^2>\varepsilon^2) \le \varepsilon^{-2}\mathbb{E}[|X_n-X|^2] \to 0$
  - Counter:
- Convergence in Probability: $X_n \xrightarrow{p} X$ in prob if $\forall \varepsilon > 0$ $\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n\to\infty} 0$
  - Pf: Omitted but not hard
  - Counter: $U \sim \text{Uniform}(0,1)$, $X_n = \begin{cases} n & U \le \frac{1}{n} \\ 0 & \text{o/w} \end{cases}$, $\mathbb{E}[X_n] = n \times \frac{1}{n} = 1$, $X_n \xrightarrow{p} X$
- Convergence in Distribution (weak convergence): $X_n \Rightarrow X$ ("weakly" / "in law") in distribution if $\mathbb{E}[h(X_n)] \xrightarrow{n\to\infty} \mathbb{E}[h(X)]$ (for every bbd cts fctn $h$)
  - $\Leftrightarrow h(X_n) \Rightarrow h(X)$

(31) **Weak Law of Large Numbers:** $X_1,\dots,X_n$ iid, mean $\mu$, var $\sigma^2$. If $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is sample mean, Then $\bar{X}_n \xrightarrow{} \mu$ in prob.

Pf: $\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu$, $\text{var}(\bar{X}_n) = \frac{1}{n^2}\text{var}(\sum X_i) = \frac{\text{var}(X_1)}{n} = \frac{\sigma^2}{n}$, pick $\varepsilon > 0$,
Chebyshev: $\mathbb{P}(|X_n - \mu| > \varepsilon) \le \frac{1}{\varepsilon^2}\frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0$

**Note:** Convergence of MGFs can force weak convergence: $Z_1, Z_2, \dots$ R.V. w/ MGF defined on $(-c,c)$, $c>0$.
If $M_{Z_n}(u) \to M_Z(u)$ for $-c<u<c$. THEN, $Z_n \Rightarrow Z$ [weak convergence]

(32) **Central Limit Theorem:** $X_1, X_2, \dots$ independent iid, mean zero, $\text{var}(X_i) = \sigma^2 < \infty$, have a common MGF defined on $(-c,c) \ni 0$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, Then $\sqrt{n}\bar{X}_n \Rightarrow N(0,\sigma^2)$

$\mathbb{E}[e^{uX}] = 1 + u\,\mathbb{E}[X] + \frac{1}{2}u^2\mathbb{E}[X^2]$ ($M(0)$, $M'(0)$)

Pf: $M(0) = \mathbb{E}[e^{0X}] = 1$, $M'(0) = 1$, $M''(0) = \sigma^2$ (moments), plug into pwr series: $M(u) = 1 + \frac{1}{2}\sigma^2u^2 + O(u^3)$
$S_n = \sum X_i$, $M_{S_n}(u) = (M(u))^n$ [iid], $Z_n = \sqrt{n}\bar{X}_n = \frac{1}{\sqrt{n}}S_n \Rightarrow M_{Z_n}(u) = M_{S_n}\left(\frac{u}{\sqrt{n}}\right) = \left(M\left(\frac{u}{\sqrt{n}}\right)\right)^n$
$M_{Z_n}(u) = \left(1 + \frac{\sigma^2u^2}{2(\sqrt{n})^2} + o\left(\frac{1}{n}\right)\right)^n \to e^{\frac{\sigma^2u^2}{2}}$ ← MGF of $N(0,\sigma^2)$, so note $\Rightarrow \sqrt{n}\bar{X}_n \xrightarrow{d} N(0,\sigma^2)$

"weak convergence preserved under continuous mappings"

(33) **Continuous Mapping Theorem:** $X_i \in \mathbb{R}^k$ R.V. $X_n \Rightarrow X$, $g: \mathbb{R}^k \to \mathbb{R}^m$ cts. Then $g(X_n) \Rightarrow g(X)$

Pf: $X_n \Rightarrow X \Rightarrow f(X_n) \Rightarrow f(X)$ $\forall$ bbd cts $f: \mathbb{R}^k \to \mathbb{R}$. Pick $h: \mathbb{R}^m \to \mathbb{R}$ bbd cts, then $h \circ g: \mathbb{R}^k \to \mathbb{R}$ bbd cts so $(X_n \Rightarrow X) \Rightarrow h(g(X_n)) \Rightarrow h(g(X))$ so ∎

(34) **Slutsky's Thm:** $X_n \Rightarrow X$, $Y_n \Rightarrow c$, then $X_n + Y_n \Rightarrow X + c$, $X_nY_n \Rightarrow cX$, $\frac{X_n}{Y_n} \Rightarrow \frac{X}{c}$ $(c \neq 0)$

Pf: omitted

# Statistical Inference

♥ of module

## Chapter 6 - Likelihood

Process:
① Get data $\{x_1, \dots, x_n\}$
② Model via $\{X_1, \dots, X_n; \theta\}$ — hypothesis tests — consistent w/ data?
③ estimate $\hat{\theta}$, accurate? — confidence intervals

(35) **Observed Likelihood function**: Data $x_1, \dots, x_n$ modelled as observed values of R.V.s w/ joint density/pmf $f_{X_1,\dots,X_n}(x_1, \dots, x_n; \theta)$, Observed likelihood function is $L(\theta) = f_{X_1,\dots,X_n}(x_1, \dots, x_n; \theta)$

- Not pmf/density so we need to integrate to 1, function of $\theta$, data is fixed!

(MLE)

(36) **Maximum Likelihood Estimate**: MLE $\hat{\theta}$ is the value of $\theta$ that maximises $L(\theta)$ assuming it exists & is unique: $\hat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$ [same ∵ log monotonic]

log-likelihood

$\ell(\theta) = \log L(\theta)$

MLE { (i) makes the data more probable; (ii) value of $\theta$ that has most support from data }

Warning: correlation (coincidence of an event happening) $\not\Rightarrow$ causation.
- Rain CAUSES ppl to stay indoors [makes sense]
- staying indoors CAUSES rain [≠ make sense]

beware of confounding variables

(37) **Example MLE Calculations**: CORE SKILL OF THE MODULE!!!

Assumption: INDEPENDENCE!!
↳ Q: How were populations sampled?
① If your sample from internet - you self-select for nerdy internet kinds...
② Phone for political survey: NOT EVERYONE PICKS UP THE PHONE ⇒ NOT a random sample of pop. of interest...
③ Snowball sampling: ask friends to ask to answer ⇒ Health/wealth Qs ⇒ very bad!

Getting responses > good quality data — dangerous!!!

### (a) MLE for sequence of Bernoulli R.V.s

only records $\sum x_i$ & $n$ so good for privacy!

- Likelihood: $L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$ $\xrightarrow{\log}$ $\ell(\theta) = \log\theta(\sum x_i)$
- Diff: $\frac{\partial \ell}{\partial \theta} = \frac{1}{\theta}\sum x_i - \frac{1}{1-\theta}(n - \sum x_i)$ $+\log(1-\theta)(n-\sum x_i)$
- $\frac{\partial \ell}{\partial \theta} = 0 \iff \dots \iff \hat{\theta}_{MLE} = \frac{1}{n}\sum x_i$
- JV: $L(\theta) \to 0$ when $\theta \to 0$ or $\theta \to 1 \Rightarrow$ MAX

### (b) MLE for data modelled by binomial RV

- Model: $Y = \sum_{i=1}^{n} X_i$ w/ $X_i$ binomial & remember $n$, $Y \sim \text{Binomial}(n, \theta)$
- Likelihood: $L(\theta) = P_\theta(Y = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \Rightarrow \ell(\theta) = \log\binom{n}{y} + y\log\theta + (n-y)\log(1-\theta)$
- Diff: $\frac{\partial \ell}{\partial \theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta}$, $\frac{\partial \ell}{\partial \theta} = 0 \iff \hat{\theta}_{MLE} = \frac{y}{n}$
- JV: $\theta \to 1$, $\theta \to 0$, $\ell(\theta) \xrightarrow{\theta\to 1} 0$, $L(\theta) \xrightarrow{\theta\to 0} 0 \Rightarrow$ MAX

### (c) MLE for data modelled by Normal RVs (I)

$-\frac{1}{2}\sigma^{-2} \to \frac{1}{2}\sigma^{-1}$

- Model $X_1, \dots, X_n \sim n$ iid $N(\mu, \sigma^2)$ $\quad \theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$
- $L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \Rightarrow \ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$ : log easier to diff
- Diff $\frac{\partial \ell}{\partial \mu} = 0$ & $\frac{\partial \ell}{\partial \sigma} = 0 \Rightarrow \frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\frac{\partial \ell}{\partial \sigma} = \frac{-n\,2\pi\sigma}{2\pi\sigma^2} - \frac{1}{2\sigma}\sum(x_i-\mu)^2$
- JV: Hessian has strictly negative eigenvalues / $L(\mu, \sigma^2) \to 0$ on boundary of $\mathbb{R}\times(0,\infty) \Rightarrow$ MAX
- So can say that $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

### (d) MLE for data modelled by discrete uniform RVs

no need for calculus here!!!

- Model: bag has unknown # balls ($n$ — goal to estimate). Select $k$ at random: data $x_1, \dots, x_k$ each #s ∧ (sample WITH replacement)
  ↳ use $X_1, \dots, X_k$ as independent Uniform($\{1, \dots, n\}$) to model this data.
- $L(n) = \prod_{i=1}^{k}\frac{1}{n}\mathbb{1}\{x_i \in \{1, \dots, n\}\} = \frac{1}{n^k}\mathbb{1}\{n > \max\{x_1, \dots, x_k\}\}$
- $L(n)$ monotone ∵ $\frac{1}{n^k} \Rightarrow L(n) \downarrow$ for $n > \max\{x_1, \dots, x_k\} \Rightarrow \hat{n}_{MLE} = \max\{x_1, \dots, x_k\}$

### (e) MLE for simple linear regression [Normal RVs (II)]

"linear regression w/ normally distributed errors"

"all models are wrong but some are useful"

calculations easier

- Model: $x_i$ = age (explanatory variable / covariate), $y_i$ = blood pressure (random, influenced by age) $\Rightarrow Y_i = \gamma + \beta x_i + \sigma\varepsilon_i$ ; $\varepsilon_i \sim$ iid $N(0,1)$ — params
- re-parametrise: $\bar{x} = \frac{1}{n}\sum x_i$, $x_i' = x_i - \bar{x}$ [centered], $Y_i = \alpha + \beta x_i' + \sigma\varepsilon_i$ $[\alpha = \gamma + \beta\bar{x}]$
- switch: $\varepsilon_i = \frac{1}{\sigma}(y_i - \alpha - \beta x_i')$ (use formula) $\Rightarrow L(\alpha, \beta, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i'))^2$
- Diff: $\frac{\partial \ell}{\partial \alpha} = 0 \iff \hat{\alpha}_{MLE} = \bar{y}$, $\frac{\partial \ell}{\partial \beta} = 0$, $\frac{\partial \ell}{\partial \sigma} = 0 \iff \hat{\beta}_{MLE} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum(y_i - (\hat{\alpha} + \hat{\beta}(x_i - \bar{x})))$
- JV: can $L(\alpha, \beta, \sigma) \to \infty$ as $(\alpha, \beta, \sigma) \to \partial(\mathbb{R}\times\mathbb{R}\times(0,\infty))$ ?
  ↳ Involved, check LN. Generalises to multiple covariates. VERY important! GLMs...

(f) MLE for multinomial Distribution

"Think political opinion poll: # individuals → categories"

- Setup: sample $n$ individuals from pop. w/ $k$ types. $n_i$ of type $i$, want $p_i$, $\sum_{i=1}^{k} n_i = n$
- Model: $N_1, \dots, N_k \sim \text{Multinomial}(n, p)$, $p \in \mathbb{R}^k$. $\sum p_i = 1$, $p_i > 0 \ \forall i \in \{1, \dots, n\}$ $p_k = 1 - p_1 - \dots - p_{k-1}$
- Likelihood: $L(p) = \binom{n}{n_1 \dots n_k} \prod_{i=1}^{k} p_i^{n_i}$, $n \in \mathbb{R}^k$, max $p$ over parameter space (Lagrange multipliers)
- Maximizer: $\ell(p) = \log\binom{n}{n_1 \dots n_k} + \sum_{i=1}^{k} n_i \log p_i$, For $i = \{1, \dots, k-1\}$, split ... $\Rightarrow \hat{p}_{i_{MLE}} = \frac{n_i}{n}$
- JU: when $n_i > 0 \ \forall i$, $L(p) \to 0$ as $p_i \to 0$

major tool in modern data science

"when we raise money its AI, when we hire its machine learning, when we do the work, it's logistic regression."

blurry line between stats, ML, data science etc... often data science sounds sexy & gets funding but substance is ...

(g) MLE for Logistic regression

- Scenario: $x_i$ = BMI, $y_i = \begin{cases} 1 & \text{type II diabetes} \\ 0 & \text{otherwise} \end{cases}$
- Model: $Y_i \sim \text{Bernoulli}(f(x_i; \beta_0, \beta_1))$ where $f(x; \beta_0, \beta_1) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$ logistic function!
- Attractive for interpretation: $1 - f(x; \beta_0, \beta_1) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)} \Rightarrow \log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)}\right) = \beta_0 + \beta_1 x$ linear
- Likelihood: $L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(x_i; \beta_0, \beta_1)^{y_i} (1 - f(x_i; \beta_0, \beta_1))^{1 - y_i}$
- MLE: compute numerically.

---

## Chapter 7 - Repeated Sampling Principal

How accurate are our MLE estimates?

Does NOT depend on the model params !!! e.g... MLE

(38) **Statistic**: A statistic is simply a function $t(x_1, \dots, x_n)$ of the observed data $x_1, \dots, x_n$.

(39) **Repeated Sampling Principal**: Statistical decision procedures should be evaluated on the basis of their behaviour in hypothetical repetitions of the experiment that generated the original data $x_1, \dots, x_n$. If we observe data that $\neq$ fit model $\Rightarrow$ reject model.

note: if data $x_1, \dots, x_n$ modelled by $X_1, \dots, X_n$ w/ param $\theta$, if $t(x_1, \dots, x_n)$ estimates model parameter $\theta$, then $T = t(X_1, \dots, X_n)$ is an estimator for $\theta$.

(40) **Bias of an Estimator**: Bias of the estimator $T = T(X_1, \dots, X_n)$ for $\theta$ is $\mathbb{E}_\theta[T - \theta]$ $= \mathbb{E}_\theta[T] - \theta$

want small BUT large +ve & -ve don't cancel...

want both small

(41) **Mean-Square-Error of an estimator**: MSE of $T$ for $\theta$ is $\mathbb{E}_\theta[(T - \theta)^2]$

distance between estimator & param

Square so ± cancel out

scales linearly rather than quadratically

(42) **Standard error of an estimator**: S.E. $\approx$ RMSE $\approx$ RMSD of $T$ for $\theta$ is $\sqrt{\mathbb{E}_\theta[(T - \theta)^2]}$

root mean squared deviation

MSE = variance + Bias²

(43) **Bias-variance decomposition of MSE**: $MSE = \mathbb{E}_\theta[(T - \theta)^2] = \text{Var}_\theta(T) + (\mathbb{E}_\theta(T) - \theta)^2$

pf: $\mathbb{E}[(T - \theta)^2] = \mathbb{E}[(T - \mathbb{E}_\theta[T] + \mathbb{E}_\theta[T] - \theta)^2]$

$= \underbrace{\mathbb{E}[(T - \mathbb{E}_\theta[T])^2]}_{\text{variance}} + 2\mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])(\mathbb{E}_\theta[T] - \theta)] + \underbrace{(\mathbb{E}_\theta[T] - \theta)^2}_{\text{bias}^2}$ (non-random)

(44) **Consistency of Estimators**: A sequence of estimators $T_1, T_2, \dots$ is consistent for $\theta$ if $T_n \xrightarrow{n \to \infty} \theta$ in probability.

note: $T_n = T_n(X_1, \dots, X_n)$ estimates $\theta$. $MSE(T_n) = \underbrace{\mathbb{E}_\theta[(T_n - \theta)^2] \xrightarrow{n \to \infty} 0}_{\text{convergence in 2-mean.}}$

(30) $\Rightarrow T_n \to \theta$ in probability

note: MLE could be biased but consistent.

- Data $x_1, \dots, x_n \sim$ iid Uniform$(0, \theta)$, $\theta$ unknown
- MLE $\hat{\theta} = M_n = \max\{x_1, \dots, x_n\}$, $L(\theta) = \frac{1}{\theta^n}$ for $\theta > \max\{x_1, \dots, x_n\}$
- CDF of $M_n$: $F_{M_n}(m) = \left(\frac{m}{\theta}\right)^n$ for $0 < m < \theta$, $\mathbb{P}(M_n \le m) = (\mathbb{P}(X_1 \le m))^n$
- PDF of $M_n$: $f_{M_n}(m) = \frac{n m^{n-1}}{\theta^n}$ for $0 < m < \theta$ [diff]
- $\mathbb{E}_\theta[M_n] = \int m \cdot \frac{n m^{n-1}}{\theta^n} dm = \frac{n}{n+1}\theta \Rightarrow$ bias $= \mathbb{E}_\theta[M_n] - \theta = -\frac{1}{n+1}\theta$
- $\mathbb{E}_\theta[M_n^2] = \dots = \frac{n\theta^2}{n+2}$, $MSE = \mathbb{E}_\theta[(M_n - \theta)^2] = \dots = \frac{2}{(n+1)(n+2)}\theta^2$

more calc shows biased but not consistent...

(45) **MLE General Picture**: Under 'suitable regularity conditions' (***), sequences of MLE estimators are [i] asymptotically unbiased, [ii] sequences of MLEs have MSE which are asymptotically as small as poss.

Hence MLE consistent

(***) { • sequences of MLEs $\neq$ converge to boundary of param space • # data $\to \infty$ • dim parameters $\theta$ cannot increase } for sequences of MLEs

MLE might not be the 'best' estimator in terms of MSE

(46) **MLE may not yield optimal MSE:** idea is that intuitively natural estimators can actually be biased. Need to adjust to make them unbiased.

$\{MSE = \text{variance} + \text{bias}^2$ Pick your poison...

Say we wanna estimate $\mu, \sigma^2$...

This is straightforward

(a) Estimation of mean using sample mean

- Take $X_1, \dots, X_n$ iid mean $\mu$, variance $\sigma^2$. If $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is sample mean
- $\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum \mathbb{E}[X_i] = \frac{n\mu}{n} = \mu \Rightarrow$ Bias: $\mathbb{E}[\bar{X}_n] - \mu = 0 \Rightarrow$ unbiased
- $MSE = \text{var} + \text{bias}^2 = \text{var}(\bar{X}_n) = \frac{1}{n^2}\sum \text{var}(X_i) = \frac{\sigma^2}{n}$ ← MSE

MLE variance is biased, sample variance is unbiased but ↑ variance

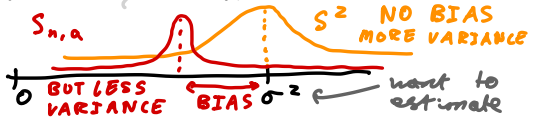
$\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \iff \mathbb{E}[Z^2] = \text{var}(Z) + \mathbb{E}[Z]^2$

(b) Estimation of variance using sample variance

- MLE is biased! $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.
- Sample variance is unbiased! $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = n\mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}_n^2]$
- Why? $\mathbb{E}[S^2] = \frac{1}{n-1}\mathbb{E}[\sum_{i=1}^n X_i^2 - 2\sum_{i=1}^n \bar{X}_n X_i + \sum_{i=1}^n \bar{X}_n^2] = \frac{1}{n-1}\left(n\mathbb{E}[X_i^2] - 2n\mathbb{E}[\bar{X}_n^2] + n\mathbb{E}[\bar{X}_n^2]\right)$
  So bias $= \mathbb{E}[S^2] - \sigma^2 = 0$ $= \frac{n}{n-1}\left([\text{var}(X_i) + \mathbb{E}[X_i]^2] - [\text{var}(\bar{X}_n) + \mathbb{E}[\bar{X}_n]^2]\right)$
- Let $S_{n,a} = \frac{1}{a}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ $= \frac{n}{n-1}\left((\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2)\right) = \frac{n-1}{n-1}\sigma^2 = \sigma^2$
- $\text{argmin}(MSE(S_{n,a}))$ $= a^* = n+1$

$MSE(\underbrace{\hat{\sigma}^2_{MLE}}_{\text{biased, better MSE}}) < MSE(\underbrace{S^2}_{\text{unbiased, worse MSE}})$

Trade off between bias & variance



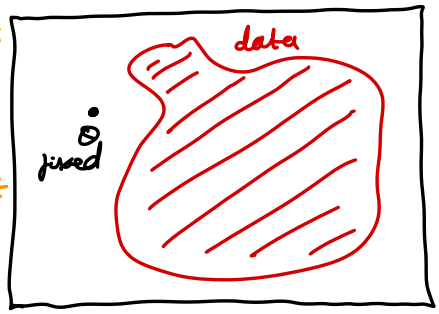

**note:** method of moments is an alternative to MLE. 200 years old. MLE is gold standard!

AIM: construct intervals using data ⇒ where unknown $\theta$ lies...

## Chapter 8 - Confidence Intervals

(47) **Confidence Interval:** Data $x_1, \dots, x_n$ modelled by $X_1, \dots, X_n$ w/ param $\theta \in \Theta$. A pair of statistics $\ell(x_1, \dots, x_n)$, $r(x_1, \dots, x_n)$ specify an interval $[\ell(x_1, \dots, x_n), r(x_1, \dots, x_n)]$ which is called a $100(1-\alpha)\%$ CI for $\phi(\theta)$ if for any $\theta \in \Theta$, $\mathbb{P}_\theta[\ell(X_1, \dots, X_n) \le \phi(\theta) \le r(X_1, \dots, X_n)] = 1 - \alpha$

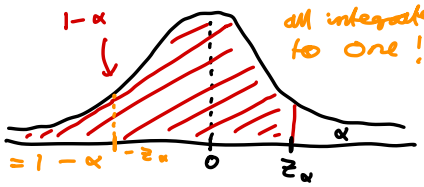
"Throw data at wall: $100(1-\alpha)\%$ chance of hitting $\theta$ in your wet sponge of data"




(48) **Gaussian percentiles:** For $\alpha \in [0,1]$ set the $100(1-\alpha)$ percentile of the standard Normal distribution to be $z_\alpha$ s.t. if $Z \sim N(0,1)$, then $\mathbb{P}(Z \le z_\alpha) = 1 - \alpha$

**note:** (a) $\mathbb{P}(Z \le z_\alpha) = 1 - \alpha \iff \Phi(z_\alpha) = 1 - \alpha \iff z_\alpha = \Phi^{-1}(1-\alpha)$

(b) $\mathbb{P}(Z > -z_\alpha) = \mathbb{P}(Z < z_\alpha) = 1 - \alpha$




(49) **Normal CI with known variance, unknown mean:** data $x_1, \dots, x_n$ modelled by $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ unknown $\mu$, KNOWN $\sigma^2 > 0$. Choose MLE estimate for $\mu$: $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ $\left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow \mathbb{P}_{\mu,\sigma^2}[a \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le b] = \Phi(b) - \Phi(a)$ has prob. $1-\alpha$ of containing $\mu$

$-a = b = z_{\alpha/2}$

(50) **Pivot:** is a random variable whose distribution $\not\equiv$ depend on params of underlying model

**note:** If we can find a pivot of simple form ⇒ easy to build a CI!!!

(51) **Student's t-distribution:** $Z \sim N(0,1)$, $V \sim \chi^2_n$ are independent R.V.s $T = \frac{Z}{\sqrt{V/n}} \sim t_n$

← think 'spread out' $N(0,1)$

$n$ degrees of freedom

(52) **More Normal confidence Intervals:** Data $x_1, \dots, x_n$ modelled by $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

(a) Estimation of variance (ignore mean)

- MLE: $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n}S^{*2}$ where $S^{*2} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ ← unbiased sample variance
- Fischer's thm ⇒ $\hat{\mu}$ independent from $\hat{\sigma}^2$ and $\frac{n-1}{\sigma^2}S^{*2} = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2 \sim \chi^2_{n-1}$ ← pivot!
- Pivot: $\alpha \in (0,1) \Rightarrow \mathbb{P}_{\mu,\sigma^2}(\chi^2_{n-1,1-\alpha/2} \le \frac{n-1}{\sigma^2}S^{*2} \le \chi^2_{n-1,\alpha/2}) = 1 - \alpha$

$\iff \mathbb{P}_{\mu,\sigma^2}\left(\frac{(n-1)S^{*2}}{\chi^2_{n-1,\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^{*2}}{\chi^2_{n-1,1-\alpha/2}}\right) = 1 - \alpha$

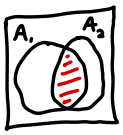
$\chi^2_{n-1,\alpha}$ is the $100(1-\alpha)$ percentile of $Y \sim \chi^2_{n-1}$: $\mathbb{P}(Y \le \chi^2_{n-1,\alpha}) = 1 - \alpha$

$Z = \frac{\bar{X}_n - M}{\sigma/\sqrt{n}} \sim N(0,1)$, $V = \frac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$

(b) Estimation of mean w/ variance unknown

- Setup: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\bar{X}_n - M}{\sigma/\sqrt{n}} \sim N(0,1)$, $S^{*2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- Fisher's thm: $\frac{n-1}{\sigma^2} S^{*2} \sim \chi^2_{n-1}$ independent of $\Rightarrow T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\bar{X}_n - M}{\sigma/\sqrt{n}} \cdot \frac{1}{\sqrt{\frac{(n-1)S^{*2}}{\sigma^2(n-1)}}} \sim t_{n-1}$
- CI: pivot is $T = \sqrt{n}\frac{\bar{X}_n - \mu}{S^*} \Rightarrow \mathbb{P}(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2})$

Note: widths of CI for $\mu$ & $\sigma^2$ are not independent...




$\mathbb{P}(A_1 \cap A_2) \ge 1 - \mathbb{P}(A_1^c) - \mathbb{P}(A_2^c)$

$\mu \in CI_\mu$, $\sigma^2 \in CI_{\sigma^2}$ — w/ prob $1 - 2\alpha$, $\mu, \sigma^2 \in$ their CI

$\Rightarrow \mathbb{P}_{\mu,\sigma^2}\left(\bar{X}_n - t_{n-1,\alpha/2}\frac{S^*}{\sqrt{n}} \le \mu \le \bar{X}_n + t_{n-1,\alpha/2}\frac{S^*}{\sqrt{n}}\right) = 1-\alpha$

$\Rightarrow \left[\bar{x} - t_{n-1,\alpha/2}\frac{s^*}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s^*}{\sqrt{n}}\right]$ is $100(1-\alpha)\%$ CI for $\mu$

(c) Regression case (use fisher's thm for linear regression)...

(S3) CI for Binomial Distribution: Untrown proportion $p$ of pop. carry genetic marker

- $\hat{p} = \frac{x}{n}$ where $x$ carry marker in sample size of $n$, model as $X = \sum_{i=1}^n B_i$, $B_i \overset{iid}{\sim}$ Bernoli(p)
- CLT: $x$: sum iid R.V. $\Rightarrow X \overset{d}{\approx} N(\underbrace{np}_{\text{mean}}, \underbrace{np(1-p)}_{\text{var}})$ (S2) $\Rightarrow \frac{(\frac{X}{n} - p)\sqrt{n}}{\sqrt{p(1-p)}} \overset{d}{\approx} N(0,1)$ [doesn't involve p]
  - approximate pivot
  - Horrible CI ∵ nasty function of p...
- WLLN: $\frac{X}{n} \xrightarrow{p} p$
- Cts mapping thm: $\sqrt{\frac{x}{n}(1-\frac{x}{n})} \xrightarrow{d} \sqrt{p(1-p)}$
- Slutsky's thm: $\frac{\sqrt{n}(\frac{x}{n} - p)}{\sqrt{\frac{x}{n}(1-\frac{x}{n})}} \xrightarrow{d} N(0,1)$
  - no p in denom

$\Rightarrow \mathbb{P}_p\left[-z_{\alpha/2} \le \frac{\sqrt{n}(\frac{x}{n} - p)}{\sqrt{\frac{x}{n}(1-\frac{x}{n})}} \le z_{\alpha/2}\right] = 1-\alpha$

$\Rightarrow \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

approximate CI for $p$ → need big $n$

(S4) Multinomial CI: lead of one political party over another

- Setup: data $n_1, \dots, n_k$; $\sum n_i = n$, modelled by multinomial$(n, p)$. Estimate $p_1 - p_2$
- MLE: $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = \frac{n_1}{n} - \frac{n_2}{n}$
- Mean: $\mathbb{E}[\hat{p}_1 - \hat{p}_2] = \dots = p_1 - p_2$. $\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \mathrm{Var}(\hat{p}_1) - 2\mathrm{cov}[\hat{p}_1, \hat{p}_2] + \mathrm{Var}(\hat{p}_2) = \dots$ (cov from covariance matrix)
- Approx pivot: $\frac{\frac{n_1 - n_2}{n} - (p_1 - p_2)}{\sqrt{\frac{(\hat{p}_1 - \hat{p}_2) - (\hat{p}_1 - \hat{p}_2)^2}{n}}} \approx N(0,1) \Rightarrow \hat{p}_1 - \hat{p}_2 \pm \sqrt{\frac{(\hat{p}_1 - \hat{p}_2) - (\hat{p}_1 - \hat{p}_2)^2}{n}} \times z_{\alpha/2}$

replace some params w/ their estimators via WLLN / CLT

approx CI for $p_1 - p_2$

(S5) Median instead of mean:

- $X_1, \dots, X_n \overset{iid}{\sim}$ cauchy $\Rightarrow \frac{1}{n}\sum X_i \sim$ cauchy $\Rightarrow$ cannot estimate $\theta$ based on $\bar{X}_n$
- Trick: compute median instead of mean! Define $Y_i = \begin{cases} 1 & \text{if } X_i \le \theta \\ 0 & \text{o/w} \end{cases}$ $Y_i \sim \mathrm{Bern}(\frac{1}{2})$
- $\mathbb{P}(X_{(k)} \le \theta \le X_{(n-k+1)}) = \mathbb{P}(\sum_{i=1}^n Y_i \in (k, n-k+1)) = A_k$
- Pick $k$ s.t. $\mathbb{P}(A_k) \ge 1-\alpha$.
- Pivot: $\sum_{i=1}^n Y_i \sim \mathrm{Binomial}(n, \frac{1}{2})$
- Have another book... a little confusing

From estimating ↦ Testing!

## Chapter 9 - Statistical Tests

(S6) Testing setup: Given data $x_1, \dots, x_n$ modelled by a joint distribution depending on $\theta \in \Theta$. The statistical model → likelihood $L(\theta; x_1, \dots, x_n)$ (depends on data). Define:

(i) null hypothesis $\theta \in \Theta_0 \subset \Theta$

(ii) alternative hypothesis $\theta \in \Theta_1 \subset \Theta$

intuition: $L(\theta) \approx$ how much does data support param value $\theta$?

possible param space

conventionally, these are nested hypotheses $\Theta_0 \subset \Theta_1$

- can be simple: $\Theta_0 = \{\theta_0\}$
- or compound: $\Theta_0 = (-\infty, \theta_0]$ ($\theta$ is idiom)

(S7) Likelihood Ratio: $W(x_1, \dots, x_n) = W(\Theta_0, \Theta_1) = \frac{\sup_{\theta \in \Theta_1}\{L(\theta)\}}{\sup_{\theta \in \Theta_0}\{L(\theta)\}}$

- best case scenario under alt. hypothesis
- best case scenario under null hypothesis

[Likelihood Ratio Test - LRT]

↗ unlikely if null hypothesis true ⇒ small $p$ ⇒ reject $\Theta_0$

measures the maximum probability of seeing evidence against the null hypothesis when null hypothesis is actually true.

Note: Large value of $W$ ⇒ "evidence vs null hypoth."

↳ There are $\theta \in \Theta_1$ that are much more strongly supported by the data than any $\theta \in \Theta_0$

(S8) P-value of a Test Statistic: Suppose $W(x_1, \dots, x_n)$ is a statistic to test $\Theta_0$ against $\Theta_1$, $W(X_1, \dots, X_n)$ is R.V.

The p value for $W(x_1, \dots, x_n)$ is

$p(x_1, \dots, x_n) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(X_1, \dots, X_n) \ge W(x_1, \dots, x_n))$

↳ small p value ⇒ strong evidence vs null.

How often will test statistic be wrong?

(59) **Decision Theoretic Interpretation:** Null hypothesis simple so $\Theta_0 = \{\theta_0\} \subset \Theta$, [nested]

ⓐ Pick significance level $\alpha \in (0,1)$ [Based on contextual evaluation of risk of making mistakes]

ⓑ Model the scenario: obtain data $x_1, \dots, x_n$ & model w/ R.V.s $\longrightarrow$ [likelihood $L(\theta; x_1, \dots, x_n)$]

ⓒ Test null hypothesis vs. alt. hypothesis using LRT (or other chosen test)

ⓓ Consider the resulting p-value:

- $p(x_1, \dots, x_n) \le \alpha \Rightarrow$ reject null hypothesis $\Theta = \Theta_0$ at significance level $\alpha$
- $p(x_1, \dots, x_n) > \alpha \Rightarrow$ do not reject $\Theta_0$ at s.l. $\alpha$ [$\not\Rightarrow$ accept $\Theta_0$ - could be wrong!]

**Useful Lemma:** $X$ R.V. w/ cts CDF $F_X$, Then $\mathbb{P}(F_X(X) \le \alpha) = \alpha \quad \alpha \in [0,1]$

(60) **Uniform Distribution of p-value under a simple null hypothesis:**

Suppose $\Theta_0 = \{\theta_0\}$, $W(X)$ has cts dist for $\theta = \theta_0$, Then, under null, $\mathbb{P}_{\theta_0}$, p-value $p(X) \sim \text{Uniform}(0,1)$

pf: Let $Z, X$ be R.V. $Z \perp X$, distributed $\mathbb{P}_{\theta_0}$. by test statistic def (distribution of $W(X)$ when $\theta = \theta_0$)

$$p(X) = \mathbb{P}_{\theta_0}(W(Z) \ge W(X) \mid X) = 1 - \mathbb{P}_{\theta_0}(W(Z) < W(X) \mid X) = 1 - F_W(W(x); \theta_0)$$

$$\mathbb{P}_{\theta_0}(p(X) \le \alpha) = \mathbb{P}_{\theta_0}(1 - F_W(W(x); \theta_0) \le \alpha) = \mathbb{P}_{\theta_0}(F_W(W(x); \theta_0) > 1-\alpha) = 1 - \mathbb{P}_{\theta_0}(F_W(W(x); \theta_0) \le 1-\alpha)$$

($1-(1-\alpha) = \alpha$)

(The t-test)

(61) **One Sample t-test:**

- Setup: data $x_1, \dots, x_n$ modeled by $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $\Theta_0 = \{(\mu,\sigma): \mu = 0, \sigma > 0\}$ (null hypothesis), $\Theta_1 = \{(\mu,\sigma): \mu \in \mathbb{R}, \sigma > 0\}$ (alternative hypothesis)
- Likelihood Ratio: $W(\Theta_0, \Theta_1) = \frac{L(\bar{x}, \hat{\sigma}^2)}{L(0, \hat{\sigma}^2)} = \dots = \left(1 + \frac{n\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{n/2} = \left(1 + \frac{t^2}{n-1}\right)^{n/2}$ where $t = \frac{\bar{x}}{\sqrt{s^2/n}}$, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
- Fischer's thm: under $\mu = 0$, $t(X) \sim t_{n-1}$ [JU] ($F_{t_{n-1}}$ is $t_{n-1}$ CDF)
- $W(x)$ $\uparrow$ function of $t^2 \Rightarrow p(x) = \mathbb{P}_{0,\sigma^2}(W(X) > W(x)) = \mathbb{P}_{0,\sigma^2}(t(X)^2 \ge t(x)^2) = 2(1 - F_{t_{n-1}}(|t(x)|))$
- Reject null hypothesis if $|t(x)| \ge t_{n-1, \alpha/2}$

be very careful what $\alpha$ value you choose!

note: rejecting null hypothesis $\iff \theta_0 \notin$ 100(1-$\alpha$)% confidence interval for $\mu$. LIWT!

(62) **Significance & power of a test:** $\Theta$ - parameter space, $\mathcal{H}_0 : \theta \in \Theta_1$ (null hypothesis), $\mathcal{H}_1 : \theta \in \Theta_2$ (alt. hypothesis), $\Theta_1 \subset \Theta_2 \subset \Theta$

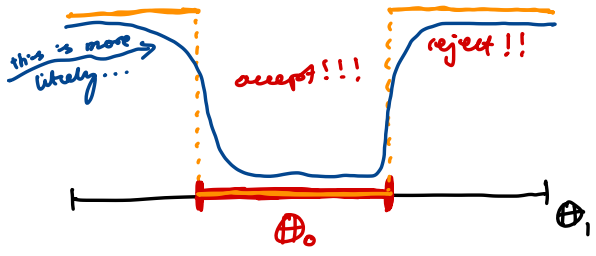
- A test has significance level $\alpha$ if it has prob $\le \alpha$ of rejecting $\mathcal{H}_0$ when it is true
- A test $\rightarrow$ test statistic $W(x)$: reject $\mathcal{H}_0$ if $W(x) \ge c$ where $c$ chosen to achieve sig. level $\alpha$. Set $R_\alpha = \{x : W(x) \ge c\}$. Reject $\mathcal{H}_0$ if $x \in R_\alpha$ [rejection region]

| | $\mathcal{H}_0$ | $\neg\mathcal{H}_0$ ← truth |
|---|---|---|
| $R$ ← reject | I | |
| $\neg R$ | | II |

- Type I error: rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is true.
  - ↳ significance/size $\alpha$ bounds $\mathbb{P}$(Type I error): $\sup_{\theta \in \Theta_0} \mathbb{P}(W(x) \ge c) = \sup_{\theta \in \Theta_0} \mathbb{P}(x \in R_\alpha) \le \alpha$
- Type II error: not rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ false.
  - ↳ The power of a test controls $\mathbb{P}$(Type II error): $\beta(\theta) = \mathbb{P}_\theta(X \notin R_\alpha)$ for $\theta \in \Theta_1 \setminus \Theta_0$
- In brief:
  - Power function of a test: $\beta(\theta) = \mathbb{P}_\theta(\mathcal{H}_0 \text{ rejected})$
  - Significance level/size of test: $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$

Power = $\mathbb{P}$(reject $\mathcal{H}_0$) when $\mathcal{H}_0$ false
Size = max $\mathbb{P}$(reject $\mathcal{H}_0$) when $\mathcal{H}_0$ true

Note: Ideally $\beta(\theta) \approx 1$ for $\theta \in \Theta_1 \setminus \Theta_0$
$\alpha \ge \beta(\theta) \approx 0$ for $\theta \in \Theta_0$
↳ Not possible ∵ $\beta(\theta)$ is continuous...



Bookwork to recall:

Ⓐ **MGFs specify Distributions:** If $X, Y$ are two random variables on $\mathbb{R}$ with $M_X(t) = M_Y(t)$ on some interval $(-c, c)$ with $0 \in (-c, c)$. Then $X$ and $Y$ have the same distributions

$X$ abs cts R.V.

Ⓑ **Univariate Transformation:** If $Y = g(X)$ where $g: \mathbb{R} \to \mathbb{R}$ is increasing, bijective, continuously differentiable ~~& positive~~, then $F_Y(y) = F_X(g^{-1}(y))$ and $f_Y(y) = \frac{1}{g'(g^{-1}(y))} f_X(g^{-1}(y))$

Ⓒ **$X \sim MVN$:** $X \in \mathbb{R}^n$ has MVN distribution if $X = \mu + AZ$ where $\mu \in \mathbb{R}^n$ is the mean vector, $Z \in \mathbb{R}^m$ is an $m$-vector of $m$ iid $N(0,1)$ random numbers and $A \in \mathbb{R}^{n \times m}$ is a matrix such that $V = AA^T \in \mathbb{R}^{n \times n}$ is the variance-covariance matrix, so that

$$X \sim MVN(\mu, V)$$

Ⓓ A distribution is memoryless if $\mathbb{P}(X > u+t \mid X > t) = \mathbb{P}(X > u)$

Ⓔ **Fischer's theorem:** If $X_1, \dots, X_n$ are a sequence of $N(\mu, \sigma^2)$ random variables, then setting $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^{*2} = \sum_{i=1}^n (X_i - \bar{X}_n)^2$, then the following are independent

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

Ⓕ **Convergence:**

- $X_n \to X$ in 2-mean if $\mathbb{E}[(X_i - X)^2] \xrightarrow{n\to\infty} 0$
- $\Downarrow$
- $X_n \to X$ in prob if $\forall \varepsilon > 0$ $\mathbb{P}(|X_n - \mu| > \varepsilon) \xrightarrow{n\to\infty} 0$
- $\Downarrow$
- $X_n \to X$ in distribution if $\mathbb{E}[h(X_n)] \xrightarrow{n\to\infty} \mathbb{E}[h(X)]$ $\forall$ bdd $h: \mathbb{R} \to \mathbb{R}$

Ⓖ **WLLN:** $X_1, X_2, \dots$ sequence of random variables iid with finite mean $\mu$ and variance $\sigma^2$. If $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean, then $\bar{X}_n \to \mu$ in prob

Pf: $\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \le \frac{\text{var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$ $\forall \varepsilon > 0$

Ⓗ **CLT:** $X_1, X_2, \dots$ sequence of iid random variables w/ zero mean and finite variance, and have a common MGF defined on some interval $(-c, c)$ with $0 \in (-c, c)$. Then $\sqrt{n}\,\bar{X}_n \Rightarrow N(0, \sigma^2)$

Ⓘ **cts mapping thm:** If $X_n \Rightarrow X$ in $\mathbb{R}^m$ and $g: \mathbb{R}^m \to \mathbb{R}^n$ is continuous then $g(X_n) \Rightarrow g(X)$

Slutsky's & cts mapping are both convergence in dist

Ⓙ **Slutsky's thm:** If $X_n \Rightarrow X$, $Y_n \Rightarrow c$, then (i) $X_n + Y_n \Rightarrow X + c$ (ii) $X_n Y_n \Rightarrow Xc$ and (iii) $\frac{X_n}{Y_n} \Rightarrow \frac{X}{c}$ if $c \neq 0$

Ⓚ An estimator is consistent if $T_n \xrightarrow{p} \theta$

Ⓛ **p-value:** $p(x_1, \dots, x_n) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(X) \ge W(x))$. Max probability of seeing evidence against the null hypothesis when $H_0$ is actually true.

Ⓜ Type I: $\mathbb{P}$(reject $H_0$ when $H_0$ true)
Type II: $\mathbb{P}$(don't reject $H_0$ when $H_0$ false)

Rejection region $R_\alpha = \{x : w(x) \geq c\}$ where $c$ chosen to achieve s.l. $\alpha$.
reject $H_0$ if $x \in R_\alpha$

$$\sup_{\theta \in \Theta_0} P_\theta(w(x) \geq c) = \sup_{\theta \in \Theta} P_\theta(x \in R_\alpha) \leq \alpha$$

Power $\beta(\theta) = P_\theta(H_0 \text{ rejected})$

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

---

Recall checks:

- If $\mathbb{E}[e^{tx}] < \infty$ $\forall t \in (-c, c)$ then MGF of $X$ $m_X(t) = \mathbb{E}[e^{tx}] = 1 + t\mathbb{E}[x] + \frac{1}{2}t^2\mathbb{E}[x^2] + \ldots$ $\forall t \in (-c, c)$ where $c > 0$ some constant.
- $\int_x^\infty \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_x^\infty = e^{-\lambda x}$ ← need?

# List of Sneaky Tricks

- $X_1, \dots, X_n$ w/ $X_i \sim$ Exponential$(\lambda)$
  - $\sum_{i=1}^{n} X_i \sim$ Gamma$(n, \lambda)$ [Sum of $n$ iid exp is gamma]
  - $\mathbb{P}(X > y) = e^{-\lambda x}$ which is easier to work w/ !
  - $\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\frac{1}{Y}\right] = \int_0^\infty \frac{1}{y} \frac{1}{\Gamma(n)} \lambda^n y^{n-1} e^{-\lambda y} dy$ $\quad Y \sim$ Gamma$(n, \lambda)$
  - Memorylessness $\mathbb{P}(X > u+t \mid X > u) = \mathbb{P}(X > t) \quad \forall u, t > 0$
  - Gamma$(n, \frac{1}{2}) \sim \chi^2_n$

- Uniform distribution on a weird area: $\frac{1}{\text{Area}(A)} \int_{\mathbb{R}^2} \mathbb{I}\{(x,y) \in A\}\, dA$
  ↳ use symmetry $\text{Cov}(X, Y) = \text{Cov}(X, -Y)$

- Transformation thm requires STRICTLY $\uparrow$ function

- State WHEN you use a named theorem.
  ↳ Sum iid Random variables $\Rightarrow$ CLT applies
  ↳ exp cts $\Rightarrow$ apply CMT

- Fisher's Thm for sequence $X_1, \dots, X_n$ but $\mu, \sigma^2$ unknown.
  ↳ $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2_{n-1}$
  so $\mathbb{E}\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] = \sigma^2 \mathbb{E}[\underbrace{Z_1^2 + \dots + Z_{n-1}^2}_{\substack{n-1 \quad Z \sim N(0,1) \text{ R.V.} \\ \text{so var} = 1}}] = (n-1)\sigma^2$

- Be VERY careful!!! NOT $X_1 \sim \text{Exp}(\underbrace{\lambda}_{\lambda_1}), X_2 \sim \text{Exp}(\underbrace{\lambda}_{\lambda_2})$

- Justify MLE maxes.

- NOTE: SHIFTED DISTRIBUTIONS $\Rightarrow$ $X = Z + 2$ (not odd factors)

- DEFINE ALL YOUR VARIABLES. (leave nothing unexplained.

- CLT
  - iid
  - common MGF

  WLLN
  - iid
  - finite $\mu, \sigma^2$.

- Say test statistic $w(x) \downarrow$, defined on $x > 0$, then $w$ will be biggest on $[0, c]$ so that is the rejection region.
  ↳ use $\uparrow$ / $\downarrow$ of test statistic to determine shape of $R_\alpha$

- $X \sim N(0,1) \quad Y = X^2$

  ↳ $F_Y(y) = P(Y \le y) = P(X^2 \le y) = \int_{-\infty}^{\infty} \mathbb{I}\{x^2 \le y\} dx = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x) dx$

  ↳ $f_X(y) = F_Y'(y)$

USE $\mathbb{I}$ functions to identify ranges

- $Z \in \mathbb{R}^n \quad \mathbb{E}[Z] = \mu$

  ↳ $V_{ij} = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])]$

  $= \mathbb{E}[Z_i Z_j] - \mu_i \mu_j \implies V = ZZ^T + \mu\mu^T$

- conditional expectation:
  - Take out what's known
  - Tower property (can create an extra $\mathbb{E}[\cdot | X]$ in!

- Sum $n$ iid R.V. ⇒ apply CLT ⇒ log N.

- Several Normals ⟶ $MVN(\mu, V)$ where $V \in \mathbb{R}^{n \times n}$

  ↳ Specify rows/columns of matrix carefully. $AA^T = V$

  ↳ watch out for $\chi^2_{n-2}$ at any point ← ALL RELATED!

- $Y_i = \gamma + \beta x_i + \sigma \varepsilon_i \implies Y_i \sim N(\gamma + \beta x_i, \sigma^2)$ ⟶ PDF!

- Lagrange multipliers...

- For $c \in I$

  ↳ $\mu, \sigma^2$ unknown ⇒ Fishers, $t_{n-1}$ create it!

  ↳ $\sigma^2$ known ⇒ simpler $\chi^2_{n-1}$.

$$T = \frac{Z}{\sqrt{\frac{V}{n}}} \sim t_n$$

$Z \sim N(0,1)$

$V \sim \chi^2_n$

- Likelihood ratio statistic ⇒ PICK values of parameters in diff scenarios. E.g. $\mu = \frac{n\bar{x} + m\bar{y}}{n+m}$

- WLLN proof: $\forall \varepsilon > 0$

$$P[|X_n - \mu| > \varepsilon) = P((X_n - \mu)^2 \ge \varepsilon^2) \le \frac{\mathrm{Var}(\bar{X}_n - \mu)}{\varepsilon^2} = \frac{n\sigma^2}{n^2\varepsilon^2} \xrightarrow{n \to \infty} 0$$

- Choose questions CAREFULLY!

  ↳ make sure you can do all parts/ have ideas for how to attack.

- If you can't do a logical step, LEAVE IT & come back.

  ↳ all double, just can't go wrong.