

# STAT 420 (Theory and Methods of Statistics)

Instructor: Joan Ren

Textbook: Introduction to Mathematical Statistics (Hogg, McKean, Craig)

Type: [#class](#) [#source](#)

Topics: [Statistics](#) [Probability](#) [Hypothesis Testing](#)

---

## Probability

- Some basic examples to introduce the **sample space**, **events**, etc.
  - Def: The **sample space** is the set of all possible outcomes of an experiment
  - Def: An **event** is a subset of the sample space
  - Ex 1: Toss a die and observe the upper face
    - Sample space:  $S = \{1, 2, 3, 4, 5, 6\}$
    - Event:  $A = \{\text{observing an even number}\} = \{2, 4, 6\}$
  - Ex 2: Measure the amount of milk in a 12 oz bottle
    - Sample space:  $S = \{x \mid 0 \leq x \leq 12\} = [0, 12]$ , where  $x$  is amount of milk in bottle
    - Event:  $A = \{\text{at most one third of the bottle}\} = \{x \mid 0 \leq x \leq 4\} = [0, 4]$
- Operations on sets; if  $A$  and  $B$  are sets, we have:
  - Def (**union**):  $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$
  - Def (**intersection**):  $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$
  - Def (**complement**):  $A^c = \{x \mid x \notin A, x \in S\}$
- Disjointness
  - Def: two events  $A$  and  $B$  are **disjoint** or **mutually exclusive** if  $A \cap B = \emptyset$
  - Def: the events  $A_1, \dots, A_n$  are **pairwise disjoint** if  $A_i \cap A_j = \emptyset$  for all  $i \neq j$
- Probability
  - Def: **probability** is a number between 0 and 1 which measures the *likelihood of occurrence of an event*
  - Back to first Ex:  $P(\{1\}) = \dots = P(\{6\}) = \frac{1}{6}$ 
    - Def: in the above case, we say the outcomes in  $S$  are **equally likely**
  - Not all sample spaces have equally likely outcomes
    - Ex: investing in a business venture
      - $S = \{s, f\}$  (success, failure)
      - $P(\{s\})$  and  $P(\{f\})$  can only be assigned/determined based on previous experience and data analysis
  - Properties of probability
    - $P(\{A^c\}) = 1 - P(\{A\})$
    - $P(\{\emptyset\}) = 0$
    - $A \subset B$  implies  $P(A) \leq P(B)$
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

## Conditional probability

- Back to first Ex: let  $A = \{2, 4, 6\}$ ,  $B = \{\text{observation} \leq 3\} = \{1, 2, 3\}$ 
  - Suppose we know  $B$  occurs; then  $P(A|B) = \frac{1}{3}$

- "Probability of  $A$  given  $B$ "

- "Given  $B$ " means we only consider the elements in  $B$

- Formula:  $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Ex: toss a fair coin twice and observe the upper face; let  $A$  be an event observing a head on the first toss and  $B$  be an event observing a tails on the second toss. Show  $P(B|A) = P(B) = \frac{1}{2}$
- Sol: Note  $S = \{HH, HT, TH, TT\}$ . If  $B$  is observed, the sample space is restricted to  $\{HT, TT\}$ . If  $A$  is observed, the sample space is restricted to  $\{HT, HH\}$ . Now, note

$$P(A) = P(\{HH\}) + P(\{HT\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(B) = P(\{HT\}) + P(\{TT\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Then, we have  $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2}$ , so  $P(B|A) = P(B)$  as desired

### • Independence

- Def: Two events  $A$  and  $B$  are independent if  $P(A \cap B) = P(A)P(B)$
- Caution: for arbitrary events  $A$  and  $B$ , we usually do not have  $P(A \cap B) = P(A)P(B)$ 
  - In Ex 1, let  $C = \{1, 3\}$ ; then,  $A = \{2, 4, 6\}$  and  $P(A \cap C) = 0$  as  $A \cap C = \emptyset$

### • Bayes' theorem

- Def: a set of events  $A_1, A_2, \dots$  is a **partition** of a sample space  $S$  if:
  - $A_1, A_2, \dots$  are pairwise disjoint
  - $\bigcup_{i=1}^{\infty} A_i = S$
- **Law of Total Probability**
  - Def: let  $B$  be any event.  $P(B) = \sum_{i=1}^{\infty} P(A_i)P(B|A_i)$
- Formula:

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{P(B)}$$

- Ex (1.4.15)
  - Sol:

$$P(WWRW) = \frac{7}{10} \cdot \frac{7}{10} \cdot \frac{3}{10} \cdot \frac{7}{10}$$

$$P(RWWW) = \frac{3}{10} \cdot \left(\frac{7}{10}\right)^3$$

$$P(WWWR) = \left(\frac{7}{10}\right)^3 \cdot \frac{3}{10}$$

$$P(WRWW) = \frac{7}{10} \cdot \frac{3}{10} \cdot \left(\frac{7}{10}\right)^2$$

$$P(\text{exactly one red ball in 4 trials}) = \sum_{i=1}^4 P(\text{exactly one red ball at position } i)$$

$$= 4 \cdot \frac{3}{10} \cdot \left(\frac{7}{10}\right)^3$$

Note: must keep 4 significant figures after decimal point in the course

- Ex (1.4.33)

- Sol: We want  $P(\text{nonsmoker}|\text{death})$ ; if  $x = P(\text{death}|\text{nonsmoker})$ , Bayes' theorem yields

$$\begin{aligned} P(\text{nonsmoker}|\text{death}) &= \frac{P(\text{death}|\text{nonsmoker})P(\text{nonsmoker})}{P(\text{death})} \\ &= \frac{x \cdot 0.65}{6x \cdot 0.10 + 3x \cdot 0.25 + x \cdot 0.65} \\ &= \frac{0.65x}{2x} = 0.325 \end{aligned}$$

## Random Variables

- Def: a **random variable** is a function that assigns a real number to each outcome in the sample space
  - Ex: toss a coin 3 times and observe the upper faces and let  $X$  be the number of heads observed
    - Note all possible values of  $X$  are  $\{0, 1, 2, 3\}$
  - Ex: record the number of TV sets sold in a store and let  $X$  be the number of sets sold
    - All possible sets of  $X$  are  $\{0, 1, \dots\}$  which is infinitely many and **countable** ( $X$  is a discrete r.v.)
  - Ex: measure the amount of milk in a 12 oz bottle and let  $X$  be an r.v. representing the amount of milk in the bottle
    - As  $X$  can take any value in  $[0, 12]$ ,  $X$  is a continuous r.v.

## Distributions

- Def: For a discrete r.v.  $X$ ,  $f_x = P(X = x)$  is called the **probability mass function** (Prof. Ren calls is a **probability density function**) (pm.f./p.d.f.)
- Def: For a continuous r.v.  $X$ , a function  $f(x) \geq 0$  satisfying

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

is called the **probability density function**.

- Def: For any r.v.  $X$ ,  $F(x) = P(X \leq x)$  is called the **cumulative distribution function** (c.d.f.) or the **distribution function** (d.f.). We have

$$F(x) = P(X \leq x) = \begin{cases} \sum_{t \leq x} f(t) & X \text{ discrete} \\ \int_{-\infty}^x f(t) dt & X \text{ continuous} \end{cases}$$

- Special discrete distributions
  - Bernoulli =  $\text{Bern}(p)$ 
    - pmf:  $f(x) = p^x(1-p)^{1-x}$ .  $x \in \{0, 1\}$
  - Binomial =  $\text{Bin}(n, p)$ 
    - $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ ,  $x = 0, 1, \dots, n$
  - Geometric =  $\text{Geom}(p)$ 
    - $f(x) = p(1-p)^x$ ,  $x = 0, 1, \dots$
  - Negative binomial
  - Hypergeometric
  - Poisson
    - $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ ,  $x = 0, 1, 2, \dots$
- Special continuous distributions
  - Gamma:  $G(\alpha, \beta)$ 
    - pdf:  $f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x \leq 0 \end{cases}$
  - Chi-squared
    - $\chi^2(r) = G\left(\frac{r}{2}, 2\right)$

- Exponential

- $G(1, \mu) = f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & x > 0 \\ 0 & x \leq 0 \end{cases}$

- $\mu$  is the mean

- Normal:  $\mathcal{N}(\mu, \sigma^2)$

- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}, x \in \mathbb{R}$

- Uniform:  $U(a, b)$

- $f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$

- Properties of distribution function

- $0 \leq F(x) \leq 1 \forall x \in \mathbb{R}$
  - $F(x)$  is nondecreasing and right-continuous
  - $F(x)$  is continuous (step) function if  $X$  is a continuous (discrete) r.v.
  - $P(a \leq X \leq b) = F(b) - F(a)$
  - $f(x) = \begin{cases} F(x) - F(x-) & X \text{ discrete} \\ F'(x) & X \text{ continuous} \end{cases}$
  - If  $A \subset \mathbb{R}$ ,

$$P(A) = \begin{cases} \sum_{x \in A} X & X \text{ discrete} \\ \sum_A f(x) dx & X \text{ continuous} \end{cases}$$

- Examples

- Ex (1.7.8)

- (a) We want to maximize  $f(x) = \left(\frac{1}{2}\right)^x$ , which is maximized at  $x = 1$
    - (b) We want to maximize  $12x^2(1-x)$ , where  $0 < x < 1$ 
      - $f'(x) = 12(2x - 3x^2) = 0$  gives critical points  $x = 0, \frac{2}{3}$
      - To find the global maximum, use first derivative test and check if  $f'(x) > 0$  or  $f'(x) < 0$  at critical points; this yields  $\frac{2}{3}$  as the mode

- Ex (1.7.12)

- (b) To find  $F(x)$ , we have

$$\int_{-\infty}^x f(t) dt = \int_{-\infty}^x \frac{1}{t^2} dt$$

Note that if  $x \leq 1$ ,  $f(t) = 0$ . Therefore, our integral becomes

$$\int_1^x \frac{1}{t^2} dt = 1 - \frac{1}{x}$$

So, our answer is

$$F(x) = \begin{cases} 1 - \frac{1}{x} & x > 1 \\ 0 & x \leq 1 \end{cases}$$

- Ex (1.7.24)

- We will first find  $F_Y(y) = P(Y \leq y) = P(\tan X \leq y)$  and take the derivative of  $F_Y(y)$  to obtain  $f_y$ . As  $|X| < \frac{\pi}{2}$ , we have  $|\tan^{-1} y| < \frac{\pi}{2}$ . Finding  $P(\tan X \leq y)$  is the same as finding  $P(X \leq \tan^{-1} y) = \int_{-\infty}^{\tan^{-1} y} f(x) dx$ . This is equal to

$$\int_{-\frac{\pi}{2}}^{\tan^{-1} y} \frac{1}{\pi} dx = \frac{\tan^{-1} y + \frac{\pi}{2}}{\pi}$$

As a result, we have  $F_Y(y) = \frac{1}{\pi} \tan^{-1} y + \frac{1}{2}$ . As  $f_Y(y) = F'_Y(y)$ , we have  $F'_Y(y) = \frac{1}{\pi} \frac{1}{1+y^2}$

# Expectation

- Definitions

- Def: Let  $u(x)$  be a real function and  $X$  be a r.v. Then the quantity,

$$\mathbb{E}(u(X)) = \begin{cases} \sum_x u(x)f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} u(x)f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

is called the **mathematical expectation** of  $u(x)$  or **expected value** of  $u(x)$

- Def: When  $u(x) = x$ , we have  $\mathbb{E}(X) = \mu$ , is the **mean** of  $X$
- Def: When  $u(x) = (x - \mu)^2$ , we have  $\mathbb{E}(X - \mu)^2 = \sigma^2$ , is the **variance** of  $X$

- Properties of expectation

- $\mathbb{E}[c] = c$  if  $c$  is a constant
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\text{Var}(aX + b) = a^2\text{Var}(X)$
- $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$
- $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$

- Moment generating functions

- Def: When  $u(x) = e^{tx}$  for a fixed  $x \in \mathbb{R}$ , we have  $\mathbb{E}(e^{tX}) = M_X(t)$  is the **moment generating function** (m.g.f.) of  $X$ 
  - Does not always exist, but provides a powerful tool to characterize a distribution function
  - $M_X(0) = 1$
  - $M_X(t)$  can help find the **moment** of  $F_X(x)$
- Thm: Let  $F_X$  and  $F_Y$  be two distribution functions (d.f.'s). If  $M_X(t) = M_Y(t)$  for all  $t$  in some neighborhood of 0, then  $F_X(u) = F_Y(u) \forall u \in \mathbb{R}$ .
- Thm: If  $M_X(t)$  exists for  $t \in (-\delta, \delta)$ ,  $\delta > 0$ , the following hold:
  - $M'_X(0) = \mathbb{E}(X) = \mu_X$
  - $M_X^{(k)} = \mathbb{E}(X^k)$  (this is the  $k$ th moment of  $X$ )
    - $\sigma_X^2 = M''_X(0) - [M'_X(0)]^2$  by the above

- Examples

- Ex: Expectation of an r.v. may not always exist Suppose r.v.  $X$  has p.d.f.

$$f(x) = \begin{cases} \frac{1}{x^2} & x > 1 \\ 0 & x \leq 1 \end{cases}$$

Then

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} xf(x) dx = \int_1^{\infty} x \left( \frac{1}{x^2} \right) dx \\ &= \int_1^{\infty} \frac{1}{x} dx = \ln x \Big|_1^{\infty} = \infty \end{aligned}$$

- Ex: Toss a coin three times and let  $X = \{\text{number of heads}\}$ ; then  $\mathbb{E}(X) = \frac{3}{2}$ 
  - The m.g.f. is

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) = e^0 f(0) + e^t f(1) + e^{2t} f(2) + e^{3t} f(3) \\ &= \frac{1}{8} + \frac{3}{8}e^t + \frac{3}{8}e^{2t} + \frac{1}{8}e^{3t} \quad \forall t \in \mathbb{R} \end{aligned}$$

- $M'_X(t) = \frac{3}{8}e^t + \frac{6}{8}e^{2t} + \frac{3}{8}e^{3t} \rightarrow M'_X(0) = 3/2 = \mu_X$
- Ex: Assume the m.g.f. of a r.v.  $X$  is given by

$$M_X(t) = 0.1 + 0.5e^t + 0.3e^{2t} + 0.1e^{5t}$$

- Find the p.d.f. of  $X$ : As  $M_X(t) = \sum_x f(x)e^{tx}$ , we have the pdf of  $x$  is as follows:  
 $\{x = 0 : 0.1, x = 1 : 0.2, x = 2 : 0.3, x = 5 : 0.1\}$
- Find the mean of  $X$ :  $\mu_X = M'_X(0) = 0.5e^0 + 0.6e^0 + 0.5e^0 = 0.5 + 0.6 + 0.5 = 1.6$
- Ex: 1.9.2
  - m.g.f of  $X$  is  $\mathbb{E}[e^{tX}] = \sum_{x=1}^{\infty} f(x)e^{tx} = \frac{1}{2}e^t + \frac{1}{4}e^{2t} + \dots$ 
    - This is a geometric series with common ratio  $\frac{e^t}{2}$ , now, we can use the geometric series formula, we have

$$\sum_{k=1}^{\infty} \left(\frac{e^t}{2}\right)^k = \frac{\frac{e^t}{2}}{1 - \frac{e^t}{2}}$$

- Thus,  $M_X(t) = \frac{e^t}{2-e^t}$  with domain  $t < \ln 2$ , as we need  $2 - e^t > 0$ , implying  $e^t < 2$  and  $t < \ln 2$
- $\mu_X = M'_X(0)$ ;  $M'_X(t) = \frac{2e^t}{(e^t-2)^2}$ , so  $M'_X(0) = 2$
- $\text{Var}(X) = M''_X(0) - 4$ ;  $M''_X(t) = -\frac{2e^t(e^t+2)}{(e^t-2)^3}$ , so  $M''_X(0) = -\frac{6}{-1} = 6$  and  $\text{Var}(X) = 6 - 2 = 4$
- Recall the Taylor/Maclaurin series of a function is given by

$$g(x) = g(0) + \frac{g'(0)}{1!} + \frac{g''(0)}{2!} + \dots = \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!} x^k$$

- Recall  $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ ,  $x \in \mathbb{R}$ . This implies

$$M_X(t) = \sum_{k=0}^{\infty} \frac{(t^2/2)^k}{k!} = \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k \cdot k!}$$

where  $t \in \mathbb{R}$

- Note that

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} \frac{M_X^{(k)}(0)}{k!} t^k \\ &= \sum_{k=0}^{\infty} \frac{M_X^{(2k)}(0)}{(2k)!} t^{2k} + \sum_{k=0}^{\infty} \frac{M_X^{(2k+1)}(0)}{(2k+1)!} t^{2k+1} \end{aligned}$$

- We know  $M_X^{(2k+1)}(0) = \mathbb{E}X^{2k+1}$  and  $\frac{M_X^{(2k)}(0)}{(2k)!} = \frac{1}{2^k k!}$ , so

$$\mathbb{E}[X^m] = \begin{cases} 0 & \text{if } m \text{ is odd} \\ \frac{m!}{2^{m/2}(\frac{m}{2})!} & \text{if } m \text{ is even} \end{cases}$$

## Markov's Inequality

- Statement: Let  $u(x)$  be a nonnegative function of an r.v.  $X$ . If  $\mathbb{E}[u(X)]$  exists, then  $\forall c > 0$ ,

$$P(u(X) \geq c) \leq \frac{\mathbb{E}(u(X))}{c}$$

## Chebyshev's Inequality

- Statement: If a r.v.  $X$  has mean  $\mu$  and variance  $\sigma^2$ , then  $\forall c > 0$ ,  $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$
- Proof: Let  $u(x) = (x - \mu)^2$ . Then,  $\forall x \in \mathbb{R}$ , since  $\mathbb{E}[u(x)] = \mathbb{E}[(X - \mu)^2] = \sigma^2$ . By Markov's inequality, we know that  $\forall c > 0$ ,

$$P(|X - \mu| \geq c) = P((X - \mu)^2 \geq c^2) = P(u(x) \geq c^2) \leq \frac{\mathbb{E}[u(x)]}{c^2}$$

where the last inequality follows from Markov's inequality

- Ex: If a r.v.  $X$  has mean  $\mu = 1$  and variance  $\sigma^2 = 1$ , then

$$\begin{aligned} P(-1 < X < 3) &= P(-2 < X - 1 < 2) \\ &= P(|X - 1| < 2) \\ &= 1 - P(|X - \mu| \geq 2) \quad \text{\textit{Chebyshev's inequality}} \\ &\geq 1 - \frac{\sigma^2}{2} = 1 - \frac{1}{4} = \frac{3}{4} \end{aligned}$$

## Multivariate Distributions

- Starter examples

- Ex: There are 3 blue, 2 yellow, and 1 red balls in a bag. A person randomly draws 2 balls from the bag. Let

$$\begin{cases} X &= \{\text{number of blue balls drawn}\} \\ Y &= \{\text{number of red balls drawn}\} \end{cases}$$

- All possible values of  $X$ : 0, 1, 2, all possible values of  $Y$ : 0, 1

- Thus,  $X$  and  $Y$  are discrete r.v.

- Def: The vector  $(X, Y)$  is called a **discrete random vector**

- Ex: Let  $X = \{\text{height of a newborn baby}\}$ ,  $Y = \{\text{weight of a newborn baby}\}$

- All possible values of  $X$ :  $(0, \infty)$ , all possible values of  $Y$ :  $(0, \infty)$

- Thus,  $X$  and  $Y$  are continuous r.v.

- Def: The vector  $(X, Y)$  is called a **continuous random vector**

- Bivariate distributions

- Def: The p.d.f. of a **bivariate discrete random vector** is given by  $f(x, y) = P(X = x, Y = y)$ , where

$$\sum_{x \in X} \sum_{y \in Y} f(x, y) = 1$$

- Def: The p.d.f. of a **bivariate continuous random vector** is given by  $p(x, y) = P(X = x, Y = y)$  that satisfies the following properties:

- $p(x, y) \geq 0 \forall x, y \in \mathbb{R}$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f(x, y) dy dx$$

- Def: The **joint distribution function** of a random vector  $(X, Y)$  (discrete or continuous) is given by

$$F(x, y) = P(X \leq x, Y \leq y) = \begin{cases} \sum_{u \leq x} \sum_{v \leq y} f(u, v) & \text{discrete random vector} \\ \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du & \text{continuous random vector} \end{cases}$$

- Def: The **marginal probability density function** for a random variable of a continuous random vector  $(X, Y)$  is defined as

$$f_X(x) = \begin{cases} \sum_y f(x, y) & \text{discrete random vector} \\ \int_y f(x, y) dy & \text{continuous random vector} \end{cases}$$

there is an analogous definition for  $f_Y(y)$

- Def: The **conditional probability density function** of  $Y$  given  $X = x$  is defined as

$$f_{Y|x}(y) = \frac{f(x, y)}{f_X(x)}$$

if  $f_X(x) \neq 0$ ; otherwise 0. There is an analogous definition of  $X$  given  $Y = y$

- Def: The **conditional expectation** of  $u(X)$  given  $Y = y$  is defined as

$$\mathbb{E}[u(X)|Y = y] = \begin{cases} \sum_x u(x)f_{X|y}(x) & \text{discrete} \\ \int_{-\infty}^{\infty} u(x)f_{X|y}(x) dx & \text{continuous} \end{cases}$$

There is an analogous definition for the conditional expectation of  $u(Y)$  given  $X = x$

- Def: The **conditional variance** of  $X$  given  $Y = y$  is defined as

There is an analogous definition for the conditional expectation of  $Y$  given  $X = x$

- Def: The **covariance** of  $X$  and  $Y$  is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$$

- Def: The **correlation coefficient** of  $X$  and  $Y$  is defined as  $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$

- Properties

- $|\rho| \leq 1$
- $\rho > 0$  implies  $X$  and  $Y$  are *positively correlated*, meaning for a sample  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$  from  $(X, Y)$ , when the value of  $X_i$  is large, the value of  $Y_i$  tends to be large. Similarly,  $\rho < 0$  implies  $X$  and  $Y$  are *negatively correlated* with each other (as  $\rho < 0$  implies  $\text{Cov}(X, Y) < 0$ )
- $|\rho| \approx 1$  implies that  $X$  and  $Y$  *approximately* have a linear relationship, meaning a sample  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$  from  $(X, Y)$  gives a linear straight line

- Moment generating function of  $(X, Y)$

- $M_{X,Y}(s, t) = \mathbb{E}[e^{sX+tY}]$
- $\mathbb{E}[X^k Y^m] = \frac{\partial^{k+m} M(s, t)}{\partial s^k \partial t^m} \Big|_{s=0, t=0}$

$$M_X(s) = \mathbb{E}[e^{sx}] = M_{X,Y}(s, 0)$$

$$M_Y(s) = \mathbb{E}[e^{tY}] = M_{X,Y}(0, t)$$

- $X$  and  $Y$  are independent if and only if  $M_{X,Y}(s, t) = M_X(s)M_Y(t)$  for any  $(s, t)$  in a neighborhood of  $(0, 0)$

- Independence

- Def:  $X$  and  $Y$  are **independent** if  $f(x, y) = f_X(x)f_Y(y) \forall x, y \in \mathbb{R}$  (denoted as  $X \perp Y$ ).
- If  $X \perp Y$ , then we have
  - $\text{Cov}(X, Y) = 0$  (**note**:  $\text{Cov}(X, Y) = 0$  *does not* imply  $X \perp Y$ )
  - $P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d)$  for any  $a, b, c, d, \in \mathbb{R}$
  - $\mathbb{E}[u(X)u(Y)] = \mathbb{E}[u(X)]\mathbb{E}[u(Y)]$

- Examples

- Ex (2.3.4)
  - To find  $\mathbb{E}[Y|X = 1]$ , we first compute

$$f_Y(y) = \begin{cases} \frac{7}{18} & Y = 1 \\ \frac{11}{18} & Y = 2 \end{cases}$$

- Then, we compute

$$\begin{aligned} f_{Y|X=1}(y) &= \frac{f(1, y)}{f_X(1)} \\ &= \frac{f(1, y)}{8/18} \end{aligned}$$

$$\text{which yields } f_{Y|X=1}(y) = \begin{cases} \frac{3}{8} & Y = 1 \\ \frac{5}{8} & Y = 2 \end{cases}$$

- So  $\mathbb{E}[Y|X = 1] = \frac{3}{8} \cdot 1 + \frac{5}{8} \cdot 2 = \frac{13}{8}$
- To find  $\text{Var}(Y|X = 1)$ , we compute  $\mathbb{E}[Y^2|X = 1]$  and compute  $\text{Var}(Y|X = 1) = \mathbb{E}[Y^2|X = 1] - (\mathbb{E}[Y|X = 1])^2 = 15/64$



- To find  $\mathbb{E}(3X - 2Y)$ , we use linearity of expectation with the marginal distribution functions (trivial)
- Ex (2.4.8)
  - To find  $P\left(\frac{1}{2} < X < 1, 0 < Y < \frac{3}{4}\right)$ , we compute

$$\begin{aligned}
 P\left(\frac{1}{2} < X < 1, 0 < Y < \frac{3}{4}\right) &= \int_{1/2}^1 \int_0^{3/4} 3x \, dy \, dx \\
 &= \int_{1/2}^1 \int_0^{\min(x, 3/4)} 3x \, dy \, dx \\
 &= \int_{1/2}^{3/4} 3x \, dy \, dx + \int_{3/4}^1 \int_0^{3/4} 3x \, dy \, dx \\
 &= \frac{101}{128}
 \end{aligned}$$

- To find  $\mathbb{E}(Y|X = x)$ , we first need to find  $f_X(x)$  and  $f_{Y|X}(y)$ :

$$\begin{aligned}
 f_X(x) &= \int_{-\infty}^{\infty} f(x, y) \, dy \\
 &= \int_0^x 3x \, dy \\
 \rightarrow f_X(x) &= \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}
 \end{aligned}$$

- So  $f_{Y|X}(y) = \frac{f(x, y)}{f_X(x)} = 3x/3x^2 = \frac{1}{x}$  if  $0 < y < x < 1$  and 0 otherwise

$$\begin{aligned}
 \mathbb{E}(Y|X = x) &= \int_{-\infty}^{\infty} f_{Y|X}(y) \, dy \\
 &= \int_0^x y \left(\frac{1}{x}\right) \, dy \\
 &= \frac{1}{x} \int_0^x y \, dy \\
 &= \frac{x}{2} \text{ if } 0 < x < 1
 \end{aligned}$$

- To find  $\text{Cov}(X, Y)$ , we first need to find  $f_Y(y)$  to find  $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

$$\begin{aligned}
 f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) \, dx \\
 &= \int_y^1 3x \, dx \\
 &= \frac{3}{2}(1 - y^2) \text{ if } 0 < y < 1
 \end{aligned}$$

- To find  $\mathbb{E}(X)$  and  $\mathbb{E}(Y)$ , we compute  $\mu_X = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_0^1 x(3x^2) \, dx = \frac{3}{4}$  and  $\mu_Y = \int_{-\infty}^{\infty} y f_Y(y) \, dy$
- To find  $\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$ , we compute  $\mathbb{E}(X^2)$ ,  $\mathbb{E}(Y^2)$  to find  $\sigma_X^2$  and  $\sigma_Y^2$  and can use the results from above to find  $\rho_{xy}$

#### • Linear combinations

- Let  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_m$  be r.v.s and let  $a_1, \dots, a_n$ , and  $b_1, \dots, b_m$  be constants
  - Let  $T = \sum_{i=1}^n a_i x_i = a_1 x_1 + \dots + a_n x_n$  and  $W = \sum_{j=1}^m b_j Y_j = b_1 Y_1 + \dots + b_m Y_m$
  - Thm:  $\mathbb{E}[T] = \sum_{i=1}^n a_i \mathbb{E}[X_i]$  if and  $\mathbb{E}[W] = \sum_{j=1}^m b_j \mathbb{E}[Y_j]$
  - $\text{Cov}(T, W) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j)$
  - $\text{Var}(T) = \text{Cov}(T, T) = \mathbb{E}(T^2) - (\mathbb{E}[T])^2 = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i < j} a_i a_j \text{Cov}(X_i, X_j)^2$
  - If  $X_1, \dots, X_n$  are independent, then  $\text{Var}(T) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$

#### • Ex (2.8.10)

- We want  $\rho_{12} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$
- We have  $\text{Var}(X + 2Y) = \text{Var}(X) + 4\text{Var}(Y) + 4\text{Cov}(X, Y)$ , so  $15 = 4 + 8 + 4\text{Cov}(X, Y)$

- So  $\text{Cov}(X, Y) = 3/4$
- This gives  $\rho_{12} = \frac{3/4}{\sqrt{4} \cdot \sqrt{2}} = 3/4\sqrt{2} = 3\sqrt{2}/16$

## Special Distributions

- Review
  - Chi-squared distribution
    - If  $X \sim \chi^2_{(r)}$ ,  $\mathbb{E}(X) = r$  and  $\text{Var}(X) = 2r$
  - Poisson distribution
    - If  $X \sim \text{Poisson}(\lambda)$ ,  $\mathbb{E}(X) = \lambda = \text{Var}(X)$
- Normal distributions
  - If r.v.  $X$  has a **normal distribution** with mean  $\mu$  and variance  $\sigma^2$ , we denote  $X \sim \mathcal{N}(\mu, \sigma^2)$
  - Properties
    - The m.g.f. of  $X$  is given by

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} e^{tx} \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \right) dx \\ &= e^{\mu t + \frac{\sigma^2}{2} t^2} \end{aligned}$$

- Bivariate normal distribution
  - A random vector has a **bivariate normal distribution**  $\mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$  if it has the joint pdf

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2}{2(1-\rho^2)}}$$

- It can be shown that the m.g.f. of  $(X, Y)$  is given by

$$M(t_1, t_2) = \mathbb{E}(e^{t_1 X_1 + t_2 X_2}) = e^{\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2(1-\rho^2)}}$$

- The above implies  $M_X(t) = M(t, 0) = \mathbb{E}(e^{tX+0\cdot Y}) = e^{\mu_1 t + \frac{\sigma_1^2}{2} t^2}$ , so  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 
  - Similarly, we have  $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- We can use the formula  $\frac{\partial^2 M}{\partial t_1 \partial t_2} \Big|_{(0,0)} = \mathbb{E}(XY)$  for a bivariate normal distribution as follows:

$$\begin{aligned} \frac{\partial^2 M}{\partial t_1 \partial t_2} M(t_1, t_2) &= \frac{\partial}{\partial t_2} \left( \frac{\partial M}{\partial t_1} \right) M(t_1, t_2) \\ &= M(t_1, t_2) (\mu_1 + \sigma_1^2 t_1 + \rho\sigma_1\sigma_2 t_2) + (\mu_2 + \sigma_2^2 t_2 + \rho\sigma_2\sigma_1 t_1) + M(t_1, t_2) \rho\sigma_2\sigma_1 \\ &\rightarrow \mathbb{E}(XY) = M(0, 0) \mu_1 + \mu_2 + M(0, 0) \rho\sigma_1\sigma_2 \\ &= \mu_1\mu_2 + \rho\sigma_1\sigma_2 \end{aligned}$$

- Using the above result, we obtain  $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mu_1\mu_2 + \rho\sigma_1\sigma_2 - \mu_1\mu_2 = \rho\sigma_1\sigma_2$
- Thm: If  $(X, Y) \sim \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , then  $X \perp Y$  if and only if  $\rho = 0$
- Note: In the general case of  $(X, Y)$  (not necessarily normally distributed), we have  $X \perp Y$  implies  $\text{Cov}(X, Y) = 0$ , but  $\text{Cov}(X, Y) = 0$  does not imply  $X \perp Y$
- Proof: To prove the forward direction, assume  $X \perp Y$ . Then, from  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ ,  $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ , we know

$$f_X(x)f_Y(y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2}{2(1-\rho^2)}}$$

As we must have  $f_X(x)f_Y(y)$ , we must have  $\rho = 0$ . To prove the backward direction, assume  $\rho = 0$ . Then, we have  $f(x, y) = f_X(x)f_Y(y)$ .

- Thm: If  $(X, Y) \sim \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , then
  - (a) The conditional distribution of  $Y$  given  $X = x$  is

$$P(Y|X = x) = \mathcal{N}\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)\right)$$

- (b) The conditional distribution of  $X$  given  $Y = y$  is symmetrically

$$P(X|Y = y) = \mathcal{N}\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$

- Examples

- Ex (3.5.1)

- (a) We first normalize  $Y$  using  $\mu_Y = 110$  and  $\sigma_Y = 10$  to get
 
$$P(106 < Y < 124) = P\left(\frac{-4}{10} < \frac{Y-110}{10} < \frac{14}{10}\right) = P\left(Z < \frac{14}{10}\right) - P\left(Z < -\frac{4}{10}\right),$$
 where  $Z \sim \mathcal{N}(0, 1)$  (recall this process is called standardization). Using a distribution table for this value, we obtain
 
$$P(106 < Y < 124) = 0.5746$$
    - (b) We first find  $P(Y|X = x)$  using the given formula. This gives

$$P(Y|X = 3.2) = \mathcal{N}\left(110 + 0.6 \frac{0.04}{10}(3.2 - 2.8), 100(1 - 0.6^2)\right)$$

This yields  $P(Y|X = 3.2) \sim \mathcal{N}(116, 64)$ . Finally, we can compute  $P(106 < Y < 124)$  as

$P\left(\frac{106-116}{8} < \frac{Y-116}{8} < \frac{124-116}{8}\right) = P(-1.25 < Z < 1) = P(Z < 1) - P(Z < -1.24)$  so our final answer is  $0.8413 - 0.1056 = 0.7357$

- Remark: Compare the results of (a) and (b) (one is conditional probability, the other is marginal probability); as the conditional probability is different from the unconditional probability, we have  $X$  and  $Y$  are not independent

- Thm: Let  $X_1, \dots, X_n$  be mutually independent normal r.v.s. Then, for constants  $a_1, \dots, a_n$ ,

$$\begin{aligned} Y &= \sum_{i=1}^n a_i X_i \\ &= a_1 X_1 + \dots + a_n X_n \sim \mathcal{N}\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right) \end{aligned}$$

where each  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$

- Proof: From previous results, we have  $\mu_Y = \sum_{i=1}^n a_i \mu_i$ ,  $\sigma_Y^2 = \text{Var}(Y) = \sum_{i=1}^n a_i^2 \sigma_i^2$ . We will find the mgf of  $Y$ , since if two distributions have the same mgf, they are the same distribution (by an earlier theorem). Note that

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{tY}) \\ &= \mathbb{E}\left(e^{t \sum_{i=1}^n a_i X_i}\right) \\ &= \mathbb{E}\left(\prod_{i=1}^n e^{a_i t x_i}\right) \\ &= \prod_{i=1}^n \mathbb{E}\left(e^{a_i t x_i}\right) \\ &= \prod_{i=1}^n M_{X_i}(a_i t) \end{aligned}$$

Since  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , we know  $M_{X_i}(t) = e^{\mu_i t + \frac{\sigma_i^2}{2} t^2}$ . Therefore, the above product becomes

$$\begin{aligned}\prod_{i=1}^n e^{ta_i\mu_i + \frac{\sigma_i^2}{2}(ta_i)^2} &= \prod_{i=1}^n e^{a_i\mu_i + \frac{a_i^2\sigma_i^2}{2}t^2} \\ &= e^{t(\sum_{i=1}^n a_i\mu_i) + \frac{\sum_{i=1}^n a_i^2\sigma_i^2}{2}t^2}\end{aligned}$$

Thus, we have the mgf of  $\mathcal{N}(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$  is equal to the mgf of  $Y$ , implying  $Y \sim \mathcal{N}(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$ , as desired.

- Thm: Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a **random sample** from  $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ . Then,  $(\bar{X}, \bar{Y}) \sim N_2\left(\mu_1, \mu_2, \frac{\sigma_1^2}{n}, \frac{\sigma_2^2}{n}, \rho\right)$ .

- Proof: We will prove the mgf of both distributions is the same. To compute the mgf of  $(\bar{X}, \bar{Y})$ , we have

$$\begin{aligned}M_{(\bar{X}, \bar{Y})}(t_1, t_2) &= \mathbb{E}(e^{t_1\bar{X} + t_2\bar{Y}}) \\ &= \mathbb{E}\left(e^{\frac{\sum_{i=1}^n t_1X_i + t_2Y_i}{n}}\right) \\ &= \prod_{i=1}^n M_{(X, Y)}\left(\frac{t_1}{n}, \frac{t_2}{n}\right)\end{aligned}$$

Doing some more algebra leads to the fact that the mgf of  $(\bar{X}, \bar{Y})$  is the same as that of  $\mathcal{N}_2\left(\mu_1, \mu_2, \frac{\sigma_1^2}{n}, \frac{\sigma_2^2}{n}, \rho\right)$ , so both distributions are the same.

- t-distribution

- If  $Z \sim \mathcal{N}(0, 1)$ ,  $V \sim \chi_{(r)}^2$  (where  $\chi_{(r)}^2$  represents a chi-squared distribution with  $r$  degrees of freedom), then

$$T = \frac{Z}{\sqrt{V/r}} \sim t_{(r)}$$

where  $t_{(r)}$  denotes the t-distribution with  $r$  degrees of freedom. The pdf of  $T$  is

$$f(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r} \Gamma\left(\frac{r}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{r}\right)^{\frac{r+1}{2}}}$$

where the gamma function is defined as below:

$$\begin{cases} \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} e^{-x} dx \\ \Gamma(1) &= 1 \\ \Gamma(n) &= (n-1)! & n \in \mathbb{N}^+ \\ \Gamma(\alpha) &= (\alpha-1)\Gamma(\alpha-1) & \alpha > 1 \end{cases}$$

- Mean/variance of t-distribution

- $\mathbb{E}(T) = 0$
- $\text{Var}(T) = \frac{r}{r-2}$ 
  - Note that

$$\begin{aligned}\text{Var}(T) &= \mathbb{E}(T^2) - \mathbb{E}(T)^2 \\ &= \mathbb{E}(T^2) \\ &= \mathbb{E}\left(\frac{Z^2}{V/r}\right) \\ &= \mathbb{E}\left(\frac{rZ^2}{V}\right) \\ &= r\mathbb{E}\left(Z^2 \cdot \frac{1}{V}\right)\end{aligned}$$

As  $Z \perp V$ , we have  $r\mathbb{E}\left(Z^2 \cdot \frac{1}{V}\right) = r\mathbb{E}(Z^2)\mathbb{E}(V^{-1})$ . Note that the mgf of the chi-squared distribution is

$$\mathbb{E}(X^k) = \frac{2^k \Gamma\left(\frac{r}{2} + k\right)}{\Gamma\left(\frac{r}{2}\right)} \text{ where } k > -\frac{r}{2}$$

Thus, we have

$$\begin{aligned}
 r\mathbb{E}(Z^2)\mathbb{E}(V^{-1}) &= r \cdot \mathbb{E}(Z^{-1}) \\
 &= r \frac{2^{-1}\Gamma\left(\frac{r}{2} - 1\right)}{\Gamma\left(\frac{r}{2}\right)} \\
 &= \frac{r}{2} \frac{\Gamma\left(\frac{r}{2} - 1\right)}{\left(\frac{r}{2} - 1\right)\Gamma\left(\frac{r}{2} - 1\right)} \\
 &= \frac{r}{r - 2}
 \end{aligned}$$

where the last equality follows from the fact that  $k > -\frac{4}{2}$ , so  $-1 > -\frac{r}{2}$  and  $r > 2$

- Note: the pdf of the t-distribution looks like a normal distribution, but they are not the same (to see if a distribution is normal, we must run a goodness-of-fit test to verify)
- F-distribution
  - If  $U \sim \chi^2_{(r_1)}$ ,  $V \sim \chi^2_{(r_2)}$  and  $U \perp V$  then

$$F = \frac{U/r_1}{V/r_2} \sim F_{r_1, r_2}$$

where  $F_{r_1, r_2}$  denotes the **F-distribution** with df =  $r_1, r_2$  (df = degrees of freedom)

- The pdf of  $F_{r_1, r_2}$  is given by

$$f(t) = \begin{cases} \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}}}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)} \frac{t^{\frac{r_1}{2}-1}}{\left(1+\frac{r_1}{r_2}t\right)^{\frac{r_1+r_2}{2}}} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Mean/variance
  - To find the mean, we have

$$\begin{aligned}
 \mathbb{E}(F) &= \mathbb{E}\left(\frac{U/r_1}{V/r_2}\right) \\
 &= \mathbb{E}\left(\frac{r_2}{r_1}UV^{-1}\right) \\
 &= \frac{r_2}{r_1}\mathbb{E}(U)\mathbb{E}(V^{-1}) \\
 &= \frac{r_2}{r_1}(r_1) \frac{2^{-1}\Gamma\left(\frac{r_2}{2} - 1\right)}{\Gamma\left(\frac{r_2}{2}\right)} \\
 &= \frac{r_2}{r_1}(r_1) \frac{2^{-1}\Gamma\left(\frac{r_2}{2} - 1\right)}{\left(\frac{r_2}{2} - 1\right)\Gamma\left(\frac{r_2}{2} - 1\right)} \\
 &= \frac{r_2}{r_2 - 2}
 \end{aligned}$$

- Finding the variance is left as an exercise
- MGF does not exist
- Remark: the mean of  $F_{r_1, r_2}$  does not depend on  $r_1$ . When  $r_2$  is very large, we know  $\mathbb{E}(F) \approx 1$
- Student's theorem
- Let  $x_1 \dots x_m$  be a random sample from  $\mathcal{N}(\mu, \sigma^2)$  with **sample mean**  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  and **sample variance**  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{x})^2$ . Then, the following hold:
  - (a)  $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - (b)  $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
  - (c)  $\bar{X} \perp s^2$
  - (d)  $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

- Proof: (a) follows from the previous theorem (use the mgf). To prove (b), we note that

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \rightarrow \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

From this, we have  $\left(\frac{X_i - \mu}{\sigma}\right) \sim \chi_{(1)}^2$  (proven on homework). This implies

$$\left(\frac{X_1 - \mu}{\sigma}\right)^2 + \dots + \left(\frac{X_n - \mu}{\sigma}\right)^2 \sim \chi_{(n)}^2$$

Since  $\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$ , it differs from the above by replacing  $\bar{X}$ , which reduces the degree of freedom by

1. The proof of (c) together with (b) requires linear algebra and is left as an exercise. To prove (d), note that (a) implies  $\bar{X} \sim \mathcal{N}(0, 1)$  and (c) implies  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \perp \frac{(n-1)s^2}{\sigma^2}$ , so by algebra, we have

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} \sim t_{n-1}$$

- Gamma distribution

- If  $X \sim G(\alpha, \beta)$  (where  $G$  is a gamma distribution), then its m.g.f. is given by

$$M_X(t) = \frac{1}{(1 - \beta t)^\alpha}$$

where  $t < \frac{1}{\beta}$

- Special cases

- 1.  $\text{Exp}(\mu) = G(1, \mu)$ , where  $M_{G(1, \mu)}(t) = \frac{1}{1 - \mu t}$ ,  $t < \mu$
- 2.  $\chi_{(r)}^2 = G\left(\frac{r}{2}, 2\right)$ ,  $M(t) = \frac{1}{(1 - 2t)^{r/2}}$ ,  $t < \frac{1}{2}$

## Statistical Inference

- Starter examples

- Ex: We are interested in finding the quality of 100,000 television sets.
- Ex: We want to know the percentage of Americans who support Elon Musk's recent behavior.
- Ex: We want to know the average lifetime of a certain brand of light bulbs.
- Note: in the above examples, it is impossible or impractical to obtain information on the entire population (in this case, statisticians take a sample from the population and draw conclusions based on these samples)

- Steps in statistics

- 1. Draw a sample
- 2. Analyze the sample
- 3. Draw conclusion(s)

- Intro defs

- Let  $X_1, \dots, X_n$  be  $n$  independent r.v.s each with d.f.  $F(x)$ . Then,  $X_1, \dots, X_n$  are said to be a **random sample** or an **independent and identically distributed (i.i.d.) sample** from  $F(x)$
- **Parameters** of a distribution are some numbers which describe certain characteristics of the distribution
- Let  $T = T(X_1, \dots, X_n)$  be a function of a random sample  $X_1, \dots, X_n$ . Then,  $T$  is called a **statistic**; any statistic is a **point estimator**
  - Examples of point estimators are sample means/variances/stdevs, as each of them is a function of a random sample  $X_1, \dots, X_n$
- We say  $\hat{\theta}$  is an **unbiased estimator** for  $\theta$  if  $\mathbb{E}[\hat{\theta}] = \theta$
- **Statistical inference** consists of **estimation** and **hypothesis testing**

- Estimation can be done via point estimators or **interval estimators**
- Hypothesis testing can be done via point estimators or likelihood
- Interval estimators
  - If  $X_1, \dots, X_n$  is a random sample from a distribution with mean  $\mu$ , then we know sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is a good point estimator
  - The use of the word "good" means a lot; intuitively,  $\bar{X}$  is a good point estimator for  $\mu$  because  $\hat{x}$  is close to  $\mu$  when the sample size  $n$  is large
    - But a point estimator does not tell us how reliably it estimates the parameter (what does "how reliable" mean here?)
    - In other words, we do not know how close  $\bar{X}$  is to  $\mu$
    - Ideally, for a small  $\delta > 0$ ,

$$P(|\bar{X} - \mu| \leq \delta) = 1 - \alpha$$

where  $0 < \alpha < 1$  is a small number

- Ex: if we have  $\delta = 0.01$  and  $\alpha = 0.05$ , the above equation becomes

$$P(|\bar{X} - \mu| \leq 0.01) = 0.95$$

In words, the probability the estimation error does not exceed 0.01 is 0.95

- Note the inequality  $|\bar{X} - \mu|$  implies  $-0.01 \leq \bar{X} - \mu \leq 0.01$ , which can be rewritten as  $\bar{X} - 0.01 \leq \mu \leq \bar{X} + 0.01$  can be written as  $P(\bar{X} - 0.01 \leq \mu \leq \bar{X} + 0.01)$ 
  - In words, the last expression means that with probability 0.95, the mean falls in an interval  $(\bar{X} - 0.01, \bar{X} + 0.01)$
  - We call the above interval a **95% confidence interval** or an **interval estimation** for  $\mu$
- Question: for a given  $\alpha > 0$ , what is the value of  $\delta > 0$  in  $P(|\bar{X} - \mu| \leq \delta) = 1 - \alpha$ 
  - If  $X_1, \dots, X_n$  is i.i.d,  $\mathcal{N}(\mu, \sigma^2)$ , then  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$  by the Student Theorem
  - Thus,

$$\begin{aligned} 1 - \alpha &= P(|\bar{X} - \mu| \leq \delta) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\delta}{\sigma/\sqrt{n}}\right) \\ &= P\left(|Z| \leq \frac{\delta}{\sigma/\sqrt{n}}\right) \end{aligned}$$

- Assuming we have access to a standard normal distribution table, let  $z_\alpha = \frac{\delta}{\sigma/\sqrt{n}}$ , so

$$\delta = z_\alpha \frac{\sigma}{\sqrt{n}}$$

- So, the  $(1 - \alpha)100\%$  confidence interval is given by  $\bar{X} \pm \delta = \bar{X} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$  where we assume  $\alpha$  is known

- Let  $X_1, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$  and let  $a_n = a_n(X_1, \dots, X_n) \leq b_n = b_n(X_1, \dots, X_n)$  be statistics based on the random sample. Then  $[a_n, b_n]$  or  $(a_n, b_n)$  is a  $(1 - \alpha)100\%$  confidence interval for  $\theta$  if  $P(a_n \leq \theta \leq b_n) = 1 - \alpha$
- Ex: 4.2.19
  - Let  $Y = \frac{2}{\beta} \sum_{i=1}^n X_i$ . Then,

$$\begin{aligned}
M_Y(t) &= \mathbb{E}(e^{tY}) \\
&= \mathbb{E}(e^{t \frac{2}{\beta} \sum_{i=1}^n X_i}) \\
&= \mathbb{E}(\prod_{i=1}^n e^{\frac{2t}{\beta} X_i}) \\
&= \prod_{i=1}^n M_{X_i} \left( \frac{2t}{\beta} \right) \\
&= \left[ M_X \left( \frac{2t}{\beta} \right) \right]^n
\end{aligned}$$

For  $t < \frac{1}{2}$ ,  $\left( \frac{1}{[1-\beta(\frac{2t}{\beta})]^3} \right)^n = \frac{1}{(1-2t)^{3n}}$ , so  $M_Y(t) = \frac{1}{(1-2t)^{3n}}$ , where  $t < \frac{1}{2}$ , so  $Y \sim \chi(6n)$ , which implies  $1 - \alpha = P(\chi_{1-\frac{\alpha}{2}, 6n}^2 \leq \frac{2n}{\beta} \leq \chi_{\frac{\alpha}{2}, 6n}^2)$

- Hypothesis testing

- Ex: Let  $\mu$  = [average family income in MD]. A person claims  $\mu = 35,000$ ; a statistician disputes the claim, thinking  $\mu > 35,000$ . In such a case, we have the following hypothesis test: The **null hypothesis** is  $H_0$  and states  $\mu = 35,000$ . The **alternative hypothesis** is  $H_1$  and states  $\mu > 35,000$ 
  - The null hypothesis is what is *claimed to be true* but disputed by someone.  $H_1$  is usually conjecture
  - The **goal** of a hypothesis test is to see if there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis
    - Suppose in the above example we take a sample and have  $\bar{X} = 35,010$ . Can we conclude  $H_1$  from this? Since a sample is randomly selected, there is a chance of making an error if we reject  $H_0$  using  $\bar{X} = 35,010$  to conclude  $H_1$ . We are certainly more confident to reject  $H_0$  using  $\bar{X} = 40,000$ . Intuitively, we have the following procedure:
      - Step 1: Choose a constant  $c$  sufficiently larger than \$35,000
      - Step 2: Compute data and compute  $\bar{X}$
      - Step 3: If  $\bar{X} \geq c$ , we reject  $H_0$  in favor of  $H_1$ . If  $\bar{X} < c$ , we conclude there is no sufficient evidence to reject  $H_0$
  - Question: how do we choose  $c$ ?
    - Based on the above testing procedure, we have the following situations:
      - Accept  $H_0$ ,  $H_0$  is true (correct)
      - Accept  $H_0$ ,  $H_0$  is false (type I error)
      - Reject  $H_0$ ,  $H_0$  is true (type II error)
      - Reject  $H_0$ ,  $H_0$  is false (correct)
    - Note:** if we do not find sufficient evidence to prove  $H_1$ , we must accept  $H_0$  (think of this as "innocent until proven guilty," accepting  $H_0$  does not necessarily mean  $H_0$  is true)
    - Let  $\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true})$  and  $\beta = P(\text{type II error}) = P(\text{accept } H_0 | H_0 \text{ false})$
    - We should choose  $c$  such that both  $\alpha$  and  $\beta$  are minimized (a small  $\alpha$  is especially crucial due to the goal of hypothesis testing)
      - $\alpha$  is called **level of significance** or **significance level** of the test
        - Usually, we use  $\alpha = 0.05$ ; this value is called the **statistical significance model** and is used for any hypothesis test unless a specific value of  $\alpha$  is given or required
      - The range  $\{\bar{X} \geq c\}$  is called the **rejection region** or **critical region** of the test; usually, we choose  $c$  such that the significance level  $\alpha$  is small, which is decided in advance (e.g. before you conduct a hypothesis test)
    - Choosing  $c$ 
      - Recall if  $X_1, \dots, X_n$  is a random sample from a population, with mean  $\mu$  and variance  $\sigma^2$ , if  $n$  is sufficiently large, then by CLT  $\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$



- Thus, using the original example, we compute  $\alpha$  as follows:

$$\begin{aligned} P(\text{reject } H_0 | H_0 \text{ true}) &= P(\bar{X} \geq c | \mu = 35,000) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{c - \mu}{\sigma/\sqrt{n}} \mid \mu = 35,000\right) \end{aligned}$$

- Let  $Z \sim \mathcal{N}(0, 1)$  be equal to  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ ; then, by CLT, the above probability is approximately  $P\left(Z \geq \frac{c - 35000}{\sigma/\sqrt{n}}\right)$ , which implies  $\frac{c - 35000}{\sigma/\sqrt{n}} = z_\alpha$  and  $c = 35000 + z_\alpha \frac{\sigma}{\sqrt{n}} \approx 35,000 + z_\alpha \frac{s}{\sqrt{n}}$  where  $s$  is the sample standard deviation
- Hence, the rejection region is given by

$$\bar{X} \geq 35,000 + z_\alpha \frac{\sigma}{\sqrt{n}} \implies \frac{\bar{X} - 35,000}{\sigma/\sqrt{n}}$$

- This implies that the **test statistic** is  $T = \frac{\bar{X} - 35,000}{\sigma/\sqrt{n}} \approx \frac{\bar{X} - 35,000}{s/\sqrt{n}}$
- We reject  $H_0$  in favor of  $H_1$  if  $T \geq z_\alpha$  with significance level  $\alpha$  (read the textbook for the other three cases in the hypothesis test)
- Generalized concept of hypothesis testing (read textbook for more information)
  - We have  $X_1, \dots, X_n$  i.i.d. from an unknown distribution  $F(x; \theta), \theta \in \Omega$
  - We have  $H_0$  that says  $\theta \in \Omega_0$  versus  $H_1$  that states  $\theta \in \Omega_1$ , where  $\Omega_0, \Omega_1 \subset \Omega$
  - The **power function** is given by  $P(\theta) = P(\text{reject } H_0 | \theta \in \Omega_1)$

## Consistency and Limiting Distributions

- Intuitively, we know the sample mean  $\bar{X}$  is close to the population mean  $\mu$ ; here, we study what "close" means
- Let  $\{Y_n\}$  be a sequence of r.v.'s and let  $Y$  be a r.v. We say  $Y_n$  **converges** in probability to  $Y$  if

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|Y_n - Y| \geq \epsilon) = 0$$

- This is denoted as  $Y_n \xrightarrow{P} Y$  as  $n \rightarrow \infty$
- Note: such a definition means if  $\hat{\theta}_n = Y_n = \hat{\theta}(X_1, \dots, X_n)$  is a point estimator for  $\theta$  based on a random sample  $X_1, \dots, X_n$  and if  $\hat{\theta}_n \xrightarrow{P} \theta$ , we know that for any  $\epsilon > 0$ , no matter how small, the limit of probability that  $|\hat{\theta}_n - \theta| \geq \epsilon$  is 0 as  $n \rightarrow \infty$ 
  - In such a case, we call the point estimator  $\hat{\theta}_n$  a **consistent estimator** if  $\hat{\theta}_n \xrightarrow{P} \theta$  as  $n \rightarrow \infty$
- Weak Law of Large Numbers (WLLN)
  - Thm (WLLN): If  $X_1, \dots, X_n$  is a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ , then  $\bar{x}_n \xrightarrow{P} \mu$  as  $n \rightarrow \infty$ 
    - Proof: Recall  $E(\bar{X}) = \mu$ ,  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ . By Chebyshev's Inequality, we have  $\forall \epsilon > 0 \quad P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$
    - As  $n \rightarrow \infty$ , the RHS goes to zero, so the probability also approaches 0, implying  $\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$ , implying  $\bar{x}_n \xrightarrow{P} \mu$
  - Note: WLLN implies that  $\bar{X}$  is a **consistent estimator** for  $\mu$
- Thm: Suppose  $X_n \xrightarrow{P} X$ ,  $Y_n \xrightarrow{P} Y$ . Then, the following hold:
  - (a)  $X_n + Y_n \xrightarrow{P} X + Y$
  - (b)  $X_n Y_n \xrightarrow{P} XY$
  - (c)  $aX_n \xrightarrow{P} aX$ , where  $a$  is a constant

- Thm (Continuity Theorem): Suppose  $X_n \xrightarrow{P} c$  (where  $c$  is a constant) and  $g(x)$  is a continuous function at  $x = c$ . Then,  $g(x_n) \xrightarrow{P} g(c)$  as  $n \rightarrow \infty$

- Examples

- Ex 1: Let  $X_1, \dots, X_n$  be a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Assume  $E(X_i^4) < \infty$ . Show that the sample variance  $s^2$  is a consistent estimator for  $\sigma^2$

- Sol: First, note it is relatively easy to prove that  $\bar{X}$  is an unbiased estimator of  $\mu$  using linearity of expectation, since  $E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$  so  $\bar{X}$  is an unbiased estimator of  $\mu$ . By WLLN, we can prove  $\bar{X} \xrightarrow{P} \mu$  as  $n \rightarrow \infty$ , so  $\bar{X}$  is a consistent estimator of  $\mu$ . To prove the sample variance is a consistent estimator of  $\sigma^2$  we have  $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - 2x_i\bar{X} + \bar{X}^2)$ . Then, we have

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - 2x_i\bar{X} + \bar{X}^2) &= \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - 2\bar{X} \sum_{i=1}^n x_i + n\bar{X}^2 \right) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \right) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{X}^2 \right) \\ &= \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \frac{n}{n-1} \bar{X}^2 \end{aligned}$$

By WLLN, we have  $\frac{1}{n} \sum_{i=1}^n x_i^2 \xrightarrow{P} E(X^2)$  and  $\bar{X} \xrightarrow{P} \mu$ . By the Continuity Theorem, we have  $\bar{X}^2 \xrightarrow{P} \mu^2$ . Now, note that  $\frac{n}{n-1} = \frac{1}{1-\frac{1}{n}} \rightarrow 1$  as  $n \rightarrow \infty$ . By a theorem,  $s^2 \xrightarrow{P} \mathbb{I}(E(X^2)) - \mathbb{I}(\mu^2) = E(X^2) - \mu^2 = \sigma^2$ , so  $s_n^2 \xrightarrow{P} \sigma^2$  and  $s_n^2$  is a consistent estimator of  $\sigma^2$ . We now ask: is  $s_n^2$  an unbiased estimator of  $\sigma^2$ ?

- Ex: Let  $X_1, \dots, X_n$  be i.i.d. r.v.s with distribution  $U(0, \theta)$ . Let  $Y_n = \max\{X_1, \dots, X_n\} = X_{(n)}$ . Show (a)  $Y_n$  is a biased estimator of  $\theta$  and (b)  $Y_n$  is a consistent estimator of  $\theta$ . (c)  $W_n = \frac{n+1}{n} Y_n$  is an unbiased estimator of  $\theta$ .

- Sol: For part (a), we want to find the pdf of  $Y_n$  using the distribution function of each  $X_i$ . We have  $F_{Y_n}(y) = P(Y_n \leq y) = P\{\max\{X_1, \dots, X_n\} \leq y\}$ . We note that by independence,  $\prod_{i=1}^n P(X_i \leq y) = [F_X(y)]^n$ , so

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = \begin{cases} 0 & x < 0 \\ \int_0^x \frac{1}{\theta} dt & 0 < x < \theta \\ 1 & x \geq \theta \end{cases}$$

So,

$$F_{Y_n}(y) = [F_X(y)]^n = \begin{cases} 0 & y \leq 0 \\ \left(\frac{y}{\theta}\right)^n & 0 < y < \theta \\ 1 & y \geq \theta \end{cases}$$

As  $f_{Y_n}(y) = F'_{Y_n}(y)$ , we have

$$f_{Y_n}(y) = \begin{cases} y^{n-1} \left(\frac{n}{\theta^n}\right) & 0 < y < \theta \\ 0 & \text{otherwise} \end{cases}$$

Now,

$$\begin{aligned}
E(Y_n) &= \int_{-\infty}^{\infty} f_{Y_n}(y) dy \\
&= \int_0^{\theta} y \frac{n}{\theta^n} y^{n-1} dy \\
&= \frac{n}{\theta^n} \int_0^{\theta} y^n dy \\
&= \frac{n}{\theta^n} \left( \frac{y^{n+1}}{n+1} \right)_0^{\theta} \\
&= \frac{n}{\theta^n} \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1} \theta \neq \theta
\end{aligned}$$

Thus,  $Y_n$  is a biased estimator for  $\theta$ . To prove (c), note  $E(W_n) = \frac{n+1}{n} E(Y_n) = \frac{n+1}{n} \frac{n}{n+1} \theta = \theta$ . To prove (b), note that from (a) we know  $E(Y_n) = \frac{n}{n+1} \theta$ . We will now find  $\text{Var}(Y_n)$  to use Chebyshev's inequality as follows:

$$\begin{aligned}
E(Y_n^2) &= \int_{-\infty}^{\infty} y^2 f_{Y_n}(y) dy \\
&= \int_0^{\theta} y^2 \frac{n}{\theta^n} y^{n-1} dy \\
&= \frac{n}{\theta^n} \int_0^{\theta} y^{n+1} dy \\
&= \frac{n}{n+2} \theta^2
\end{aligned}$$

So,  $\text{Var}(Y_n) = E(Y_n^2) - [E(Y_n)]^2 = \frac{n}{(n+1)^2(n+2)} \theta^2$ . By Chebyshev's inequality, we have  $\forall \epsilon > 0$ ,

$$\begin{aligned}
P(|Y_n - E(Y_n)| \geq \epsilon) &\leq \frac{\text{Var}(Y_n)}{\epsilon^2} \\
&= \frac{n\theta^2}{\epsilon^2(n+2)(n+1)^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty
\end{aligned}$$

Therefore,  $Y_n - E(Y_n) = Y_n - \frac{n}{n+1} \theta \xrightarrow{P} 0$  as  $n \rightarrow \infty$ . Also,  $\frac{n}{n+1} \theta \rightarrow \theta$  as  $n \rightarrow \infty$ , so by linearity, we have

$$E(Y_n) = E((Y_n - E(Y_n)) + E(Y_n)) = \left( Y_n - \frac{n}{n+1} \theta \right) + \left( \frac{n}{n+1} \theta \right) \xrightarrow{P} 0 + \theta = \theta$$

Thus,  $Y_n \xrightarrow{P} \theta$ , implying  $Y_n$  is a consistent estimator for  $\theta$

- Convergence in distribution

- Motivation: Sometimes we do not know the distribution of a point estimator  $\hat{\theta}_n$  but we know its approximated distribution when  $n$  is large

- Ex: Let  $X_n, \dots, X_n$  be a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . We know  $\bar{X}$  is an unbiased estimator for  $\mu$  and  $\bar{X}$  is a consistent estimator of  $\mu$ . But we do not know the exact distribution of  $\bar{X}$ . We know when  $n$  is large, by CLT, the sample mean  $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$  and
- Let  $\{X_n\}$  be a sequence of random variables and let  $X$  be a random variable. We say  $X_n$  **converges in distribution** to  $X$  if for every point  $x_0$  of  $F_X(x)$  we have  $\lim_{n \rightarrow \infty} F_{X_n}(x_0) = F_X(x_0)$  denoted by  $X_n \xrightarrow{D} X$  as  $n \rightarrow \infty$ .  $F_X$  is called the **limiting distribution** of  $X_n$
- Thm: Let  $\{X_n\}$  be a sequence of random variables with mgf  $M_{X_n}(t)$  that exists for  $-h < t < h$  for all  $n$ . Let  $X$  be a random variable with mgf  $M(t)$  which exists for  $|t| \leq h_1 \leq h$ . If  $\lim_{n \rightarrow \infty} M_{X_n}(t) = M(t)$  for  $|t| \leq h_1$ , then  $X_n \xrightarrow{D} X$
- Thm: If  $X_n \xrightarrow{P} X$ , then  $X_n \xrightarrow{D} X$  (the converse does not hold)
- Thm:  $X_n \xrightarrow{P} c$  if and only if  $X_n \xrightarrow{D} c$  where  $c$  is a constant
- Thm: If  $X_n \xrightarrow{D} X$ ,  $Y_n \xrightarrow{P} 0$ , then  $X_n + Y_n \xrightarrow{D} X$
- Thm (Continuity Theorem): If  $X_n \xrightarrow{D} X$  and  $g(x)$  is a continuous function on the support of  $X$ , then  $g(X_n) \rightarrow g(X)$  as  $n \rightarrow \infty$

- Lemma (Slutsky's Lemma): If  $X_n \xrightarrow{D} X$ ,  $A_n \xrightarrow{P} a$ ,  $B_n \xrightarrow{P} b$  where  $a, b \in \mathbb{R}$ , then  $A_n X_n + B_n \xrightarrow{D} aX + b$  as  $n \rightarrow \infty$
- Useful limits:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{b}{n}\right)^{cn} = e^{bc}$$

$$\lim_{n \rightarrow \infty} \left(1 + \frac{b}{n} + \frac{\psi(n)}{n}\right)^{cn} = e^{bc} \text{ if } \lim_{n \rightarrow \infty} \psi(n) = 0$$

- Continuing example from above

- We know  $F_{Y_n}(y) = \begin{cases} 0 & y \leq 0 \\ \left(\frac{y}{\theta}\right)^n & 0 < y < \theta \\ 1 & y \geq \theta \end{cases}$ . Then,  $\forall y \in (0, \theta)$ , we have  $\lim_{n \rightarrow \infty} F_{Y_n}(y) = \lim_{n \rightarrow \infty} \left(\frac{y}{\theta}\right)^n = 0$ . Let r.v.

$Y$  have the following distribution function:

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ 1 & y \geq \theta \end{cases}$$

This is called a **degenerate distribution function**. As  $\lim_{n \rightarrow \infty} F_{Y_n}(y) = \begin{cases} 0 & y < 0 \\ 0 & 0 < y < \theta \\ 1 & y \geq \theta \end{cases}$

Thus,  $Y_n \xrightarrow{D} Y$  as  $n \rightarrow \infty$

- Examples
- Ex: Let  $X_1, \dots, X_n$  be i.i.d.  $\mathcal{N}(0, 1)$ . Then,  $Y_n = \bar{X}_n \sim \mathcal{N}(0, 1/n)$ , so the pdf of  $Y_n$  is  $f_n(y) = \frac{1}{\sqrt{2\pi/n}} e^{-\frac{y^2}{2/n}}$  for  $y \in \mathbb{R}$ .

Does  $Y_n$  converge in distribution?

- Sol: The d.f. of  $Y_n$  is given by  $F_{Y_n}(y) = P(Y_n \leq y) = \int_{-\infty}^y f_n(t) dt$  and is computed as follows:

$$\int_{-\infty}^y f_n(t) dt = \int_{-\infty}^y \sqrt{\frac{n}{2\pi}} e^{-\frac{nt^2}{2}} dt$$

Performing a  $u$ -substitution with  $u = \sqrt{n}t$ , we obtain

$$\int_{-\infty}^{\sqrt{n}y} \sqrt{\frac{n}{2\pi}} e^{-u^2/2} du = \int_{-\infty}^{\sqrt{n}y} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

$$= \Phi(\sqrt{n}y)$$

So  $F_n(y) = \Phi(\sqrt{n}y)$  and as  $n \rightarrow \infty$ ,  $\Phi(\sqrt{n}y)$  becomes

$$G(y) = \begin{cases} \Phi(-\infty) = 0 & y < 0 \\ \frac{1}{2} & y = 0 \\ \Phi(\infty) = 1 & y > 0 \end{cases}$$

So, we have

$$\lim_{n \rightarrow \infty} F_n(y) = G(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{2} & y = 0 \\ 1 & y > 0 \end{cases}$$

Now, let r.v.  $Y$  have a degenerate d.f. with  $P(y = 0) = 1$ . Since  $Y = 0$  is not a continuous point of  $Y$ , we have  $\lim_{n \rightarrow \infty} F_n(y) = F_Y(y) \forall y \neq 0$ . Thus,  $Y_n = \bar{X}_n \xrightarrow{D} Y$  as  $n \rightarrow \infty$

- Ex: Let  $X_1, \dots, X_n \sim U(0, \theta)$  be i.i.d and  $Y_n = \max\{X_1, \dots, X_n\}$ . Note  $Y_n \xrightarrow{D} Y$  as  $n \rightarrow \infty$  if

$$F_Y(y) = \begin{cases} 0 & y < \theta \\ 1 & y \geq \theta \end{cases}$$

Consider  $Z_n = n(\theta - Y_n)$ . Does  $Z_n$  converge in distribution?

- Sol: The d.f. of  $Y_n$  is given by

$$F_{Y_n}(y) = \begin{cases} 0 & y \leq 0 \\ (y/\theta)^n & 0 < y < \theta \\ 1 & y \geq \theta \end{cases}$$

The d.f. of  $Z_n$  is given by

$$\begin{aligned} G(z) &= P(Z_n \leq z) \\ &= P(n(\theta - Y_n) \leq z) \\ &= P\left(Y_n \geq \theta - \frac{z}{n}\right) \\ &= 1 - F_{Y_n}\left(\theta - \frac{z}{n}\right) \\ &= \begin{cases} 1 & \theta - \frac{z}{n} \leq 0 \\ 1 - \left(\frac{\theta - \frac{z}{n}}{\theta}\right)^n & 0 < \theta - \frac{z}{n} < \theta \\ 0 & \theta - \frac{z}{n} \geq \theta \end{cases} \\ &= \begin{cases} 0 & z \leq 0 \\ 1 - \left(1 - \frac{z}{n\theta}\right)^n & 0 < z < n\theta \\ 1 & z \geq n\theta \end{cases} \end{aligned}$$

To prove convergence in distribution, we want  $\forall z_0 \in \mathbb{R}, \exists N_0 \in \mathbb{Z}^+$  such that  $z_0 < n\theta$  for  $n \geq N_0$ . Thus, if  $z_0 > 0$ , we have  $0 < z_0 < n\theta \forall n \geq N_0$ . Taking the limit as  $n \rightarrow \infty$ , we have

$$\lim_{n \rightarrow \infty} G_{Z_n}(z_0) = \begin{cases} 0 & z_0 \leq 0 \\ \lim_{n \rightarrow \infty} \left(1 - \left(1 - \frac{z_0}{n\theta}\right)^n\right) & 0 < z_0 < n\theta \end{cases}$$

Computing the limit gives  $1 - e^{-z_0/\theta}$ , so the final answer becomes

$$\lim_{n \rightarrow \infty} G_{Z_n}(z_0) = \begin{cases} 0 & z \leq 0 \\ 1 - e^{-z/\theta} & z > 0 \end{cases} = F_{\text{Exp}(\theta)}(y)$$

This is the  $\text{Exp}(\theta)$  cdf, so we have  $Z_n \xrightarrow{D} \text{Exp}(\theta)$  as  $n \rightarrow \infty$

- Ex: Let  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  be i.i.d.. Recall the Student's Theorem, which states  $T_n = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2/n}} \sim t_{(n-1)}$ . Show  $T_n \xrightarrow{D} \mathcal{N}(0, 1)$

- Sol: By the Student's Theorem,  $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$  and  $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{(n-1)}^2$ . Note we previously showed  $s^2 \xrightarrow{P} \sigma^2$  as  $n \rightarrow \infty$ . By the Continuity Theorem,  $s_n \xrightarrow{P} \sigma$  since  $G(x) = \sqrt{x}$  is continuous on  $(0, \infty)$ . By Slutsky's Theorem, we know the following:

$$\begin{cases} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \\ \frac{\sigma}{s_n} \xrightarrow{P} 1 \end{cases}$$

Thus,

$$T_n = \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{s_n} \xrightarrow{P} \mathcal{N}(0, 1) \cdot 1 = \mathcal{N}(0, 1)$$

as  $n \rightarrow \infty$ , so  $T_n \xrightarrow{D} \mathcal{N}(0, 1)$ , as desired.

- Ex: Let  $Y_n \sim \text{Bin}(n, p)$ . Does  $Y_n$  converge in distribution? Assume  $\mu = np$  remains the same since  $\mu$  comes from a large population with a fixed mean.
- Solution: Note  $E(Y_n) = np$ ,  $\text{Var}(Y_n) = np(1-p)$ . Also note

$$\begin{aligned} M_{Y_n}(t) &= (pe^t + 1 - p)^n \\ &= \left(1 - \frac{\mu}{n} + \frac{\mu}{n}e^t\right)^n \\ &= \left(1 + \frac{\mu}{n}(e^t - 1)\right)^n \end{aligned}$$

Taking the limit as  $n \rightarrow \infty$ , we have

$$\begin{aligned}\lim_{n \rightarrow \infty} M_{Y_n}(t) &= \lim_{n \rightarrow \infty} \left(1 + \frac{\mu}{n}(e^t - 1)\right)^n \\ &= \lim_{n \rightarrow \infty} \left[ \left(1 + \frac{1}{\frac{n}{\mu(e^t - 1)}}\right)^{\frac{n}{\mu(e^t - 1)}} \right]^{\mu(e^t - 1)} \\ &= e^{\mu(e^t - 1)} \\ &= M_Y(t)\end{aligned}$$

where  $Y \sim \text{Poisson}(\mu)$ . Thus,  $Y_n \xrightarrow{D} \text{Poisson}(\mu)$

- Ex: Let  $Z_n \sim \chi^2_{(n)}$  and  $Y_n = \frac{Z_n - n}{\sqrt{2n}}$ . Find the limiting distribution of  $Y_n$
- Sol: Recall  $M_{Z(n)}(t) = \frac{1}{(1-2t)^{n/2}}$  for  $t < \frac{1}{2}$ . Now, note that

$$\begin{aligned}M_{Y_n}(t) &= E(e^{tY_n}) \\ &= E\left(e^{t \frac{Z_n - n}{\sqrt{2n}}}\right) \\ &= E\left(e^{\frac{t}{\sqrt{2n}} Z_n - t\sqrt{\frac{n}{2}}}\right) \\ &= e^{-t\sqrt{\frac{n}{2}}} E\left(e^{\frac{t}{\sqrt{2n}} Z_n}\right) \\ &= e^{-t\sqrt{\frac{n}{2}}} M_{Z_n}\left(\frac{t}{\sqrt{2n}}\right)\end{aligned}$$

When  $\frac{t}{\sqrt{2n}} < \frac{1}{2}$ , the above expression becomes

$$e^{-t\sqrt{\frac{n}{2}}} \frac{1}{\left(1 - 2\left(\frac{t}{\sqrt{2n}}\right)^{n/2}\right)} = e^{-t\sqrt{\frac{n}{2}}} \left(1 - t\sqrt{\frac{n}{2}}\right)^{-n/2}, t < \sqrt{\frac{n}{2}}$$

Now, we have to take the limit of this expression as  $n \rightarrow \infty$ . By Taylor's Expansion of order 3, which is  $g(t) = g(t_0) + g'(t_0)(t - t_0) + \frac{g''(t_0)}{2}(t - t_0)^2 + \frac{g'''(\xi)}{3!}(t - t_0)^3$  where  $\xi$  is between  $t_0$  and  $t$ , the above expression becomes

$$e^{-t\sqrt{\frac{n}{2}}} = 1 + t\sqrt{\frac{2}{n}} + \frac{1}{2} \left(t\sqrt{\frac{2}{n}}\right)^2 + \frac{e^{\xi_n}}{6} \left(t\sqrt{\frac{2}{n}}\right)^3$$

So, because

$$M_{Y_n}(t) = \left[ e^{-t\sqrt{\frac{n}{2}}} \left(1 - t\sqrt{\frac{2}{n}}\right) \right]^{-\frac{n}{2}}$$

by the Taylor expansion we have

$$M_{Y_n}(t) = \left[ \left(1 + t\sqrt{\frac{2}{n}} + \frac{1}{2} \left(t\sqrt{\frac{2}{n}}\right)^2 + \frac{e^{\xi_n}}{6} \left(t\sqrt{\frac{2}{n}}\right)^3\right) \left(1 - t\sqrt{\frac{2}{n}}\right) \right]^{-\frac{n}{2}}$$

Now, let  $a_n = -\frac{t^3\sqrt{n}}{n^{3/2}} + \frac{\sqrt{n}e^{\xi_n}t^3}{3n^{3/2}} \left(1 - t\sqrt{\frac{2}{n}}\right) = o\left(\frac{1}{n}\right)$ . Then,  $M_{Y_n}(t)$  can be written as

$\left(1 - \frac{2}{n}t^2 + \frac{t^2}{n} + a_n\right)^{-n/2} = \left(1 - \frac{t^2}{n}\right)^{-n/2}$ . Now, we can take the limit as  $n \rightarrow \infty$  to obtain

$$\lim_{n \rightarrow \infty} M_{Y_n}(t) = \lim_{n \rightarrow \infty} \left[ \left(1 + \frac{1}{-n/t^2}\right)^{-\frac{n}{t^2}} \right]^{\frac{t^2}{2}} = e^{t^2/2}, \text{ which is the mgf of } \mathcal{N}(0, 1). \text{ Thus, } Y_n \xrightarrow{D} \mathcal{N}(0, 1) \text{ as } n \rightarrow \infty$$

- Central Limit Theorem

- Thm: Let  $X_1, \dots, X_n$  be a random sample with mean  $\mu$  and variance  $\sigma^2 > 0$ . Then,  $Y_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$  as  $n \rightarrow \infty$

- Proof: Note that  $M_{Y_n}(t) = E(e^{tY_n}) = E\left(e^{\frac{t\sqrt{n}}{\sigma}(\bar{X}-\mu)}\right)$ . This can be written as

$$E\left(e^{\frac{t\sqrt{n}}{\sigma} \frac{\sum_{i=1}^n (x_i - \mu)}{n}}\right) = E\left(e^{\frac{t}{\sigma\sqrt{n}} \sum_{i=1}^n (x_i - \mu)}\right)$$

Let  $W_i = X_i - \mu$ . Then, with  $W_i$ , the above expression becomes  $E\left(\prod_{i=1}^n e^{\frac{t}{\sigma\sqrt{n}} W_i}\right)$ . By independence of the  $W_i$  terms, this becomes  $\prod_{i=1}^n E\left(e^{\frac{t}{\sigma\sqrt{n}} W_i}\right) = \prod_{i=1}^n M_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right)$ . As the  $W_i$  are i.i.d, this becomes  $\left(M_W\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$ . From Taylor's expansion, we have  $M_W(t) = E(e^{tW}) = M_W(0) + M'_W(0)t + \frac{M''_W(\xi_t)}{2}t^2$  where  $\xi_t$  is between 0 and  $t$ . From  $M_W(0) = 1$ ,  $M'_W(0) = E(W) = 0$ , we have  $M_W(t) = 1 + \frac{M''_W(\xi_t)}{2}t^2$ . Thus,

$$M_{Y_n}(t) = \left(1 + \frac{M''_W(\xi_n)}{2} \left(\frac{t}{\sigma\sqrt{n}}\right)^2\right)^n = \left(1 + \frac{M''_W(\xi_n)t^2}{2}\right)^n$$

which can be rewritten as

$$\left[ \left(1 + \frac{1}{\frac{2\sigma^2 n}{M''_W(\xi_n)t^2}}\right)^{\frac{2\sigma^2 n}{M''_W(\xi_n)t^2}} \right]^{\frac{M''_W(\xi_n)t^2}{2\sigma^2}}$$

where  $\xi_n$  is between 0 and  $\frac{t}{\sigma\sqrt{n}}$ . Since  $\frac{t}{\sigma\sqrt{n}} \rightarrow 0$  and  $\xi_n \rightarrow 0$  as  $n \rightarrow \infty$ , we have

$$\lim_{n \rightarrow \infty} M''_W(\xi_n) = M''_W(0) = E(W^2) = E(X - \mu)^2 = \sigma^2$$

Hence,  $\lim_{n \rightarrow \infty} M_{Y_n}(t) = e^{t^2/2}$ , so  $Y_n \xrightarrow{D} \mathcal{N}(0, 1)$  as  $n \rightarrow \infty$ , as desired.

- Properties:
  - Let  $Z \sim \mathcal{N}(0, 1)$ . By the Continuity Theorem,

$$\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \xrightarrow{D} Z^2 = \chi^2_{(1)} = n \frac{(\bar{X} - \mu)^2}{\sigma^2}$$

## Maximum Likelihood Methods

- Let  $X_1, \dots, X_n \sim f(x; \theta_0)$ , where  $\theta_0 \in \Theta$  be i.i.d, where  $\Theta$  is the parameter space for  $\theta_0$  and  $f(x; \theta_0)$  is known except the value of  $\theta_0$ . This is called a **parametric model**, since there is a parameter  $\theta_0$ 
  - How do we estimate  $\theta_0$  based on  $X_1, \dots, X_n$ ?

## Maximum Likelihood Estimation

- The **maximum likelihood estimator** (MLE) is a method that can be used to estimate  $\theta_0$
- Likelihood functions
  - **Likelihood** is the probability that we observe the sample in hand which is given by

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$$

If  $f(x; \theta_0)$  is discrete, this becomes  $\prod_{i=1}^n f(x_i; \theta_0)$  and the likelihood function is  $L(\theta) = \prod_{i=1}^n f(X_i; \theta)$

- If  $X$  is discrete,  $f(x; \theta_0) = P(X = x)$
- If  $X$  is continuous, the analogous form  $P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$  can be approximated via

$$\prod_{i=1}^n P(x_i - \delta < X_i < x_i + \delta) = \prod_{i=1}^n (F_X(x_i + \delta) - F_X(x_i - \delta))$$

for a small  $\delta > 0$ . If  $\xi_i$  is between  $x_i \pm \delta$ , the above becomes

$$\prod_{i=1}^n [F'_X(\xi_i)(2\delta)] = (2\delta)^n \prod_{i=1}^n f(\xi_i; \theta_0) \stackrel{\text{small } \delta}{\approx} (2\delta)^n \prod_{i=1}^n f(x_i; \theta_0)$$

Therefore, the likelihood  $P(X_1 = x_1, \dots, X_n = x_n) \propto \prod_{i=1}^n f(x_i; \theta_0)$

- Thus, maximizing likelihood is equivalent to maximizing  $\prod_{i=1}^n f(x_i; \theta_0)$ 
  - Hence, for discrete and continuous random variables  $X$ , the likelihood function for  $\theta$  is given by  $L(\theta) = \prod_{i=1}^n f(X_i; \theta_0)$

- Maximum likelihood estimators

- The MLE  $\hat{\theta}$  for  $\theta_0$  is given by  $L(\hat{\theta}_n) = \max_{\theta \in \Theta} L(\theta)$
- What is the meaning of the MLE  $\hat{\theta}$  of  $\theta_0$ ?
  - We most likely observe the sample we have in hand if  $\theta_0 = \hat{\theta}$ . In other words, if  $\theta_0 = \hat{\theta}_n$ , the (approximate) probability of observing the sample in hand is maximized

- Important theorems

- Thm (Invariance Theorem): Under regularity (normal) conditions, if  $\hat{\theta}$  is the MLE for  $\theta$ , then  $\hat{\eta} = g(\hat{\theta})$  is the MLE for  $\eta = g(\theta)$
- Thm (Consistency Theorem): Under regularity (normal) conditions,  $\hat{\theta} \rightarrow \theta$  as  $n \rightarrow \infty$

- Examples

- Ex: Let  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$  be i.i.d. Find the MLE for  $\mu$ 
  - Sol: The likelihood function for  $\mu$  is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(X_i - \mu)^2/2} \end{aligned}$$

If  $l(\mu) = \log L(\mu) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2$ . Since  $y = \log x$  is strictly increasing, we have that if  $l(\hat{\mu}) = \max_{\mu \in \mathbb{R}} l(\mu)$ , then  $L(\hat{\mu}) = \max_{\mu \in \mathbb{R}} L(\mu)$ . So, we have

$$\begin{aligned} l'(\mu) &= -\frac{1}{2} \sum_{i=1}^n 2(X_i - \mu)(-1) \\ &= \sum_{i=1}^n (X_i - \mu) = n(\bar{X} - \mu) \end{aligned}$$

So,  $l'(\mu)$  is positive if  $\mu < \bar{X}$  and  $l'(\mu)$  is negative if  $\mu > \bar{X}$ . This means  $l(\mu)$  is increasing in the interval  $(-\infty, \bar{X})$  and decreasing in the interval  $(\bar{X}, \infty)$ , so it attains its maximum value at  $\mu = \bar{X}$  and  $\bar{X}$  is the MLE for  $\mu$

- Ex: Let  $X_1, \dots, X_n \sim U(0, \theta)$  be i.i.d. Find the MLE for  $\theta$ 
  - Sol: Note

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} & 0 < x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

So, the likelihood function for  $\mu$  is given by



$$\begin{aligned}
L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\
&= \prod_{i=1}^n \frac{1}{\theta} I\{0 < X_i \leq \theta\} \\
&= \theta^{-n} \prod_{i=1}^n I\{0 < X_i \leq \theta\} \\
&= \theta^{-n} \prod_{i=1}^n I\{X_i > 0\} I\{X_i \leq \theta\} \\
&= \theta^{-n} \left( \prod_{i=1}^n I\{X_i > 0\} \right) \left( \prod_{i=1}^n I\{X_i \leq \theta\} \right)
\end{aligned}$$

where  $I$  is an indicator random variable. let  $X_{(1)} = \min(X_1, \dots, X_n)$  and  $X_{(n)} = \max(X_1, \dots, X_n)$ . So, the above expression becomes  $\theta^{-n} I\{X_{(1)} > 0\} I\{X_{(n)} \leq \theta\}$ . As  $\theta > 0$ ,  $I\{X_{(1)} > 0\} = 1$  and

$$\theta^{-n} I\{X_{(n)} \leq \theta\} = \begin{cases} \theta^{-n} & X_{(n)} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

We can now compute the log-likelihood  $l(\theta)$

$$l(\theta) = \begin{cases} -n \log \theta & X_{(n)} \leq \theta \\ -\infty & \text{otherwise} \end{cases}$$

Takin the derivative of this, we have that  $l'(\theta) = \frac{-n}{\theta}$  when  $X_{(n)} \leq \theta$  and 0 otherwise, so  $l(\theta)$  is decreasing on the interval  $[X_{(n)}, \infty)$ , so  $l(\theta)$  attains its maximum value at  $\theta = X_{(n)}$ , so the MLE for  $\theta$  is given by

$$\hat{\theta} = X_{(n)} = \max(X_1, \dots, X_n)$$

- Ex: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d}}{\sim} \text{Poisson}(\mu)$ . Find the MLE for  $\mu$
- Sol: Since

$$f(x; \mu) = \begin{cases} \frac{\mu^x e^{-\mu}}{x!} & x \in \mathbb{Z}^{\geq 0} \\ 0 & \text{otherwise} \end{cases}$$

Since the support of  $f(x; \mu)$  does not depend on  $\mu$ , the likelihood function is given by

$$\begin{aligned}
L(\mu) &= \prod_{i=1}^n f(X_i; \mu) \\
&= \prod_{i=1}^n \frac{\mu^{X_i} e^{-\mu}}{X_i!} \\
&= e^{-n\mu} \frac{\mu^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!}
\end{aligned}$$

So the log-likelihood  $l(\mu)$  is

$$\begin{aligned}
l(\mu) &= \log L(\mu) \\
&= -n\mu + \left( \sum_{i=1}^n X_i \right) \log \mu - \sum_{i=1}^n \log(X_i!) \\
\rightarrow l'(\mu) &= -n + \frac{\sum_{i=1}^n X_i}{\mu} \\
&= \frac{n}{\mu} (\bar{X} - \mu) \\
&= \begin{cases} > 0 & \mu < \bar{X} \\ < 0 & \mu > \bar{X} \end{cases}
\end{aligned}$$

This means that  $l(\mu)$  is increasing on  $(-\infty, \bar{X})$  and decreasing on  $(\bar{X}, \infty)$ , so  $l(\mu)$  attains its maximum value at  $\mu = \bar{X}$  and the MLE for  $\mu$  is  $\bar{X}$

- Ex: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d}}{\sim} \text{Exp}(\mu)$ . Find the MLE for  $\mu$  and  $\sigma^2 = \text{Var}(X_i)$

- Sol: Note

$$f(x; \mu) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function for  $\mu$  is then

$$\begin{aligned} L(\mu) &= \prod_{i=1}^n f(X_i; \mu) \\ &= \prod_{i=1}^n \frac{1}{\mu} e^{-X_i/\mu} \\ &= \mu^{-n} e^{-1/\mu \sum_{i=1}^n X_i} \end{aligned}$$

So, the log-likelihood  $l(\mu)$  is

$$\begin{aligned} l(\mu) &= \log L(\mu) \\ &= -n \log \mu - \frac{1}{\mu} \sum_{i=1}^n X_i \\ \rightarrow l'(\mu) &= -\frac{n}{\mu} + \frac{1}{\mu^2} \sum_{i=1}^n X_i \\ &= \frac{n}{\mu^2} (\bar{X} - \mu) \\ &= \begin{cases} > 0 & \mu < \bar{X} \\ < 0 & \mu > \bar{X} \end{cases} \end{aligned}$$

This means  $l(\mu)$  is increasing on  $(0, \bar{X})$  and decreasing on  $(\bar{X}, \infty)$ . Thus,  $l(\mu)$  attains its maximum value at  $\mu = \bar{X}$ , so the MLE for  $\mu$  is given by  $\hat{\mu} = \bar{X}$ .

To find the MLE for  $\sigma^2$ , we note  $\sigma^2 = \mu^2$ , so by the Invariance Theorem, the MLE for  $\sigma^2$  is given by  $\hat{\sigma}^2 = \hat{\mu}^2 = \bar{X}^2$

- Ex: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ . Find the MLE for  $\mu$  and  $\sigma^2$ 
  - Sol: Let  $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^{\geq 0}$ , where  $\Theta$  is the parameter space. We want to find the MLE  $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$  for  $\theta = (\mu, \sigma^2)$ . The likelihood function for  $\theta$  is given by

$$\begin{aligned} L(\theta) &= L(\mu, \sigma^2) \\ &= \prod_{i=1}^n f(X_i; \theta) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i - \mu)^2 / 2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2} \end{aligned}$$

So, the log-likelihood  $l(\theta)$  is given by

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \\ &= -\frac{n}{2} \log(2\sigma) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \end{aligned}$$

As this is a 2-dimensional problem, we can find  $\frac{\partial l(\theta)}{\partial \mu}$  to find the MLE for  $\mu$  as follows:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \mu} &= -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(X_i - \mu)(-1) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i) - n\mu \\
&= \frac{n}{\sigma^2} (\bar{X} - \mu) \\
&\begin{cases} > 0 & \mu < \bar{X} \\ < 0 & \mu > \bar{X} \end{cases}
\end{aligned}$$

So, for any fixed  $\sigma^2 > 0$ ,  $l(\theta) = l(\mu, \sigma^2)$  attains its maximum value at  $\mu = \bar{X}$ , so the MLE of  $\mu$  is given by  $\mu = \bar{X}$ . To find the MLE for  $\sigma^2$ , we follow a similar process to above:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2 \\
&= \frac{n}{2\sigma^4} \left( \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \sigma^2 \right) \\
&\begin{cases} > 0 & \sigma^2 < \sigma_\mu^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \\ < 0 & \sigma^2 > \sigma_\mu^2 \end{cases}
\end{aligned}$$

This means that for fixed  $\mu$ ,  $l(\theta)$  attains its maximum value at  $\sigma^2 = \sigma_\mu^2$ . So, the MLE  $\max_\theta l(\theta)$  is equal to

$$\begin{aligned}
\max_{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+} l(\theta) &= \max_{\sigma^2 \in \mathbb{R}^+} \left( \max_{\mu \in \mathbb{R}} l(\mu, \theta) \right) \\
&= \max_{\sigma^2 \in \mathbb{R}^2} l(\bar{X}, \sigma^2) \\
&= l(\bar{X}, \sigma_\mu^2)
\end{aligned}$$

Thus, the MLE for  $\mu$  and  $\sigma^2$  are given by  $\hat{\mu} = \bar{X}$  and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ . Note: if we maximized  $\sigma^2$  first and computed  $\max_{\mu \in \mathbb{R}} (\max_{\sigma^2 \in \mathbb{R}^+})$ , this problem would be much more difficult! Also, note that  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  is *not* the sample variance equal to  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , so this estimate is *biased*. Note that

$$\begin{aligned}
E(\hat{\sigma}^2) &= E\left(\frac{n-1}{n} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\
&= \frac{n-1}{n} E(s^2) \\
&= \left(1 - \frac{1}{n}\right) \sigma^2
\end{aligned}$$

Thus, in this case the MLE  $\hat{\sigma}^2$  for  $\sigma^2$  is a biased estimator. But since  $s^2 \xrightarrow{P} \sigma^2$  i.e. sample variance is a consistent estimator of  $\sigma^2$ , we know  $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$  as  $n \rightarrow \infty$ . Hence, the MLE  $\hat{\sigma}^2$  in this example is a biased but consistent estimator for  $\sigma^2$ , since it is not equal to  $\sigma^2$  in expectation but converges to  $\sigma^2$  as  $n \rightarrow \infty$ . Note: by WLLN,  $\bar{X} \xrightarrow{P} \mu$  as  $n \rightarrow \infty$ . By the Continuity Theorem,  $\hat{\sigma}^2 = \bar{X}^2 \xrightarrow{P} \mu^2 = \sigma^2$  as  $n \rightarrow \infty$ . Thus, the MLE  $\hat{\mu} = \bar{X}$  and  $\hat{\sigma}^2 = \bar{X}^2$  are *consistent estimators* for  $\mu$  and  $\sigma^2$ , respectively

## Rao-Cramer Lower Bound and Efficiency

- Motivation: (a) What is a "good" estimator for  $\theta$ ? (b) What is the best point estimator for  $\theta$ ?
- We know that if a point estimator  $\hat{\theta}$  for  $\theta$  is an **unbiased estimator** if  $E(\hat{\theta}) = \theta$ 
  - This is desirable because it means that on average it is more likely that a point estimator is close to  $\theta$  by definition of expected value
  - If we have two point estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$  for  $\theta$ , both are unbiased estimator, which one is better?

- Since  $E(\hat{\theta}_1) = \theta$  and  $E(\hat{\theta}_2) = \theta$ , we have

$$\begin{cases} \text{Var}(\hat{\theta}_1) &= E[\hat{\theta}_1 - E(\hat{\theta}_1)]^2 = E[(\hat{\theta}_1 - \theta)^2] \\ \text{Var}(\hat{\theta}_2) &= E[(\hat{\theta}_2 - \theta)^2] \end{cases}$$

- Since  $E[(\hat{\theta}_i - \theta)^2]$  measures how much  $\hat{\theta}_i$  deviates from  $\theta$  on average, thus  $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ , then  $\hat{\theta}_1$  is a better estimator than  $\hat{\theta}_2$  for  $\theta$
- In general, a point estimator  $\hat{\theta}$  is considered to be good if  $E(\hat{\theta}) = \theta$  and  $\text{Var}(\hat{\theta})$  is small
- Among all unbiased estimators for  $\theta$ , is it possible to find an estimator that has the smallest variance? To answer this question, we consider a few definitions
- Defns

- Let  $f(x; \theta)$  be the p.d.f. of a continuous or discrete r.v.  $X$ . Its **Fisher information** is given by

$$I(\theta) = E \left[ \left( \frac{\partial \log f(X; \theta)}{\partial \theta} \right)^2 \right]$$

- Under some regularity conditions, we can show  $I(\theta) = -E \left[ \frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} \right]$

- Rao-Cramer inequality

- Thm (**Rao-Cramer inequality**): Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(x; \theta)$ ,  $\theta \in \Omega$ . If  $Y = u(X_1, \dots, X_n)$  is an unbiased estimator of  $\theta$ , then  $\sigma_Y^2 = \text{Var}(Y) \geq \frac{1}{nI(\theta)}$  where  $\frac{1}{nI(\theta)}$  is called the **Rao-Cramer lower bound** (we shorten this to R-C lower bound)

- Def: The **efficiency** of  $Y$  is defined as  $\frac{\text{R-C lower bound}}{\sigma_Y^2} = \frac{\frac{1}{nI(\theta)}}{\text{Var}(Y)}$

- Def: Let  $Y$  be an unbiased estimator of  $\theta$ . Then,  $Y$  is called an **efficient estimator** if and only if  $\sigma_Y^2$  attains the R-C lower bound, e.g.  $\sigma_Y^2 = \frac{1}{nI(\theta)}$

- Thm: If  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(X_i; \theta_0)$  and  $\hat{\theta}$  is the MLE of  $\theta_0$ , then under regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$$

- This implies  $(\hat{\theta} - \theta_0) \stackrel{D}{\approx} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right) = \mathcal{N}\left(0, \frac{1}{nI(\theta)}\right)$ , which implies  $\hat{\theta} \stackrel{D}{\approx} \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$

- This means the following:

- (a)  $E(\hat{\theta}) = \theta_0$ , so the MLE  $\hat{\theta}$  is asymptotically unbiased under regularity conditions
- (b)  $\text{Var}(\hat{\theta}) \approx \frac{1}{nI(\theta_0)}$ , which is equal to the R-C lower bound, so the MLE is asymptotically efficient

- Asymptotic efficiency

- Def: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(x; \theta_0)$ . If  $\hat{\theta}_{1n} = (X_1, \dots, X_n)$  is an estimator of  $\theta_0$  such that  $\sqrt{n}(\hat{\theta}_{1n} - \theta_0) \xrightarrow{D} \mathcal{N}(0, \sigma_1^2)$ , then we have the following:

- (a) The **asymptotic efficiency** of  $\hat{\theta}_{1n}$  is defined by  $e(\hat{\theta}_{1n}) = \frac{\frac{1}{I(\theta_0)}}{\sigma_1^2}$

- (b)  $\hat{\theta}_{1n}$  is asymptotically efficient if  $e(\hat{\theta}_{1n}) = 1$

- (c) If  $\hat{\theta}_{2n}$  is another estimator of  $\theta_0$  such that  $\sqrt{n}(\hat{\theta}_{2n} - \theta_0) \xrightarrow{D} \mathcal{N}(0, \sigma_2^2)$ , then the **asymptotic relative efficiency** (ARE) of  $\hat{\theta}_{1n}$  to  $\hat{\theta}_{2n}$  is given by  $e(\hat{\theta}_{1n}, \hat{\theta}_{2n}) = \frac{\sigma_2^2}{\sigma_1^2}$

- If  $e(\hat{\theta}_{1n}, \hat{\theta}_{2n}) > 1$ , we say  $\hat{\theta}_{1n}$  is more efficient than  $\hat{\theta}_{2n}$ 
  - In this case, we will choose to use  $\hat{\theta}_{1n}$  in practice over  $\hat{\theta}_{2n}$

- Examples

- Ex: If  $X_1, \dots, X_n$  is a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Let  $\hat{\theta}_1 = x_1$ ,  $\hat{\theta}_2 = \bar{X}$ . Then  $E(X_1) = \mu$ ,  $E(\bar{X}) = \mu$ . But

$$\begin{cases} \text{Var}(\hat{\theta}_1) &= \text{Var}(X_1) = \sigma^2 \\ \text{Var}(\hat{\theta}_2) &= \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \end{cases}$$

Since  $\text{Var}(\hat{\theta}_2) < \text{Var}(\hat{\theta}_1)$  for  $n \geq 2$ , we know  $\bar{X} = \hat{\theta}_2$  is a better point estimator for  $\mu$  than  $\hat{\theta}_1 = X_1$

- Ex: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ . In an earlier example, we showed the MLE for  $\theta = \sigma^2$  is  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ . By the Student's Thm, we know  $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{(n-1)}^2$ . This implies

$$E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = n - 1$$

and  $\text{Var}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = 2(n-1)$ . So, we have  $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2 \neq \sigma^2$ , but  $E(\hat{\sigma}^2) \rightarrow \sigma^2$  as  $n \rightarrow \infty$  by WLLN, so the MLE  $\hat{\sigma}^2$  of  $\sigma^2$  is not an unbiased estimator. Adjusting this, we have an unbiased estimator of  $\sigma^2$ :

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \text{ which is the sample variance and } E(s^2) = \sigma^2$$

- What is the efficiency of  $\hat{\sigma}^2$ ?
  - Trick question, since the MLE  $\hat{\sigma}^2$  is not an unbiased estimator and an efficient estimator must be unbiased
- What is the efficiency of  $s^2$ ?
  - We know  $E(s^2) = \sigma^2$ . By the Student's Thm,  $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{(n-1)}^2$ , so we know  $\text{Var}\left(\frac{(n-1)s^2}{\sigma^2}\right) = 2(n-1)$ , which implies

$$\frac{(n-1)^2}{\sigma^4} \text{Var}(s^2) = 2(n-1) \rightarrow \text{Var}(s^2) = \frac{2\sigma^4}{n-1}$$

To find the R-C lower bound, compute  $I(\theta)$  as follows: note  $\theta = \sigma^2 > 0$ . Then,

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\mu)^2}{2\theta}}$$

for  $x \in \mathbb{R}$ . Thus,  $\log f(x; \theta)$  is equal to

$$-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{(x-\mu)^2}{2\theta}$$

So,

$$\frac{\partial \log f(x; \theta)}{\partial \theta} = -\frac{1}{2\theta} + \frac{(x-\mu)^2}{2\theta^2}$$

and

$$\frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} = \frac{1}{2\theta^2} - \frac{(x-\mu)^2}{\theta^3}$$

so

$$\begin{aligned} I(\theta) &= -E\left(\frac{\partial^2 \log f(X; \theta)}{2\theta^2}\right) \\ &= -E\left(\frac{1}{2\theta^2} - \frac{(X-\mu)^2}{\theta^3}\right) \\ &= -\frac{1}{2\theta^2} + \frac{1}{\theta^3} E(X-\mu)^2 \\ &= -\frac{1}{2\theta^2} + \frac{1}{\theta^3} \theta \\ &= -\frac{1}{2\theta^2} + \frac{1}{\theta^2} \end{aligned}$$

so  $I(\theta) = \frac{1}{2\theta^2}$ . This implies the R-C lower bound is  $\frac{1}{nI(\theta)} = \frac{2\theta^2}{n}$ , which implies the efficiency of  $s^2$  is

$\frac{\text{R-C lower bound}}{\text{Var}(s^2)}$ , which is  $\frac{\frac{2\theta^2}{n}}{\frac{2}{n-1}\theta^2} = \frac{n-1}{n} = \left(1 - \frac{1}{n}\right) < 1$ , so  $s^2$  is not an efficient estimator of  $\theta = \sigma^2$ . On the other hand, efficiency of  $s^2 = \left(1 - \frac{1}{n}\right) \rightarrow 1$  as  $n \rightarrow \infty$ . In such a case, we say  $s^2$  is **asymptotically efficient** and the MLE  $\hat{\sigma}^2$  is an **asymptotically unbiased estimator**

- Remark: from the above example, we see the following points:

- (a) The MLE may not be an unbiased estimator, but sometimes it leads us to finding an unbiased estimator
  - (b) Sometimes, an unbiased estimator is not an efficient estimator (like sample variance) but its efficiency converges to 1 as the sample size  $n \rightarrow \infty$
  - (c) Sometimes, a biased estimator may be...
- Ex (6.2.7)
    - (a) We have

$$f(x; \theta) = \begin{cases} \frac{1}{\Gamma(4)\theta^4} x^3 e^{-x/\theta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Since  $\Gamma(4) = 3! = 6$ , we have

$$f(x; \theta) = \begin{cases} \frac{1}{6\theta^4} x^3 e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Thus,  $\log f(x; \theta) = \log 6 - 4 \log \theta + 3 \log x - \frac{x}{\theta}$ , so

$$\begin{aligned} \frac{\partial \log f(x; \theta)}{\partial \theta} &= -\frac{4}{\theta} + \frac{x}{\theta^2} \\ \rightarrow \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} &= \frac{4}{\theta^2} - \frac{2x}{\theta^3} \\ \rightarrow I(\theta) &= E\left(\frac{\partial^2 \log f(x; \theta)}{\partial \theta^2}\right) \\ &= -E\left(-\frac{4}{\theta^2} - \frac{2X}{\theta^3}\right) \\ &= -\frac{4}{\theta} + \frac{2E(X)}{\theta^3} \\ &= -\frac{4}{\theta^2} + \frac{2(4\theta)}{\theta^3} \\ &= -\frac{4}{\theta^2} + \frac{8}{\theta^2} = \frac{4}{\theta^2} \end{aligned}$$

(b) First, we have to find the MLE  $\hat{\theta}$  for  $\theta$ . We have

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\ &= \prod_{i=1}^n \frac{1}{6\theta^4} X_i^3 e^{-X_i/\theta} \\ &= (6\theta^4)^{-n} \prod_{i=1}^n X_i^3 e^{-X_i/\theta} \\ \rightarrow l(\theta) &= \log L(\theta) \\ &= -n \log 6 - 4n \log \theta + \sum_{i=1}^n 3 \log X_i - \frac{1}{\theta} \sum_{i=1}^n X_i \\ \rightarrow l'(\theta) &= -\frac{4n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n X_i \\ &= \frac{n}{\theta^2} (\bar{X} - 4\theta) \\ \rightarrow l'(\theta) &= \begin{cases} > 0 & 0 < \theta < \frac{\bar{X}}{4} \\ < 0 & \theta > \frac{\bar{X}}{4} \end{cases} \end{aligned}$$

This means  $l(\theta)$  is increasing on the interval  $(0, \frac{\bar{X}}{4})$  and decreasing on the interval  $(\frac{\bar{X}}{4}, \infty)$ , so so  $l(\theta)$  attains its maximum value at  $\theta = \frac{\bar{X}}{4}$ . Hence, the MLE for  $\theta$  is given by  $\hat{\theta} = \frac{\bar{X}}{4}$ . Now, since we want to show  $\hat{\theta}$

is an efficient estimator, we must first show that it is unbiased. We have

$$\begin{aligned}
 E(\hat{\theta}) &= E\left(\frac{\bar{X}}{4}\right) \\
 &= \frac{1}{4}E(\bar{X}) \\
 &= \frac{\mu}{4} = \frac{4\theta}{4} = \theta \\
 \text{Var}(\hat{\theta}) &= \frac{1}{16}\text{Var}(\bar{X}) \\
 &= \frac{\text{Var}(\bar{X})}{16} \\
 &= \frac{4\theta^2}{16n} = \frac{\theta^2}{4n}
 \end{aligned}$$

From (a), we have that the R-C lower bound is  $\frac{1}{nI(\theta)} = \frac{1}{n\left(\frac{4}{\theta^2}\right)} = \frac{\theta^2}{4n}$  so  $\text{Var}(\hat{\theta})$  is equal to the R-C lower bound and the MLE  $\hat{\theta}$  is an efficient estimator for  $\theta$

## Maximum Likelihood Tests

- Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(x; \theta), \theta \in \Omega$ . Consider the hypothesis test  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ . Then, the **likelihood ratio test**  $\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$  where  $\hat{\theta}$  is the MLE for  $\theta$ . We reject  $H_0$  in favor of  $H_1$  if  $\Lambda \leq c$  for some constant  $0 < c < 1$
- Our initial hypotheses are always  $H_0 : \theta \in \Omega_0$  and  $H_1 : \theta \in \Omega_1$ . The likelihood ratio test then becomes

$$\Lambda = \frac{\sup_{\theta \in \Omega_0} L(\hat{\theta})}{\sup_{\theta \in \Omega} L(\hat{\theta})} \stackrel{H_0: \theta = \theta_0, H_1: \theta \neq \theta_0}{=} \frac{L(\theta_0)}{\sup_{\theta \in \Omega_1} L(\hat{\theta})} = \frac{L(\theta_0)}{L_1(\hat{\theta})}$$

where the final equality follows by property of the MLE

- Some intuition for the rejection region  $\Lambda \leq c$  is as follows:
  - (a) Since  $L(\hat{\theta}) = \max_{\theta \in \Omega} L(\theta)$ , then we know  $L(\theta) \leq L(\hat{\theta}) \forall \theta \in \Omega$ . Thus,  $L(\theta_0) \leq L(\hat{\theta})$ , so  $0 \leq \Lambda \leq 1 \forall \theta_0 \in \Omega$
  - (b) If  $\theta_0$  is the true parameter, then  $\hat{\theta} \xrightarrow{P} \theta_0$  as  $n \rightarrow \infty$ , implying  $\hat{\theta} \approx \theta_0$  for large  $n$ . This also implies  $\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})} \approx 1$  for large  $n$
  - (c) If  $\theta_0$  is not true, then  $\hat{\theta} \not\xrightarrow{P} \theta_0$ , which implies  $\hat{\theta} \not\approx \theta_0$  for large  $n$ , which implies  $L(\theta_0) \not\approx L(\hat{\theta})$  and  $\Lambda < 1$  for large  $n$
  - In summary, we have  $\Lambda$  is large when  $H_0$  is true and  $\Lambda$  is small when  $H_0$  is not true
    - Hence, we reject  $H_0$  if  $\Lambda$  is small
- How do we choose  $c$ ?
  - Thm (Wilke's Thm): Under some regularity conditions, we know that under  $H_0$ ,  $-2 \log \Lambda \xrightarrow{D} \chi^2_{(1)}$  as  $n \rightarrow \infty$
  - In practice, we choose  $c$  for a given  $0 < \alpha < 1$  such that

$$\begin{aligned}
 \alpha &= P(\text{Type I error}) = P(\text{reject } H_0 | H_0) \\
 &= P(\Lambda \leq c | \theta = \theta_0) \\
 &= P(-2 \log \Lambda \leq -2 \log c | \theta = \theta_0) \\
 &\stackrel{\text{Wilke's Thm}}{\approx} P\left(\chi^2_{(1)} \geq 2 \log c\right) \\
 &\rightarrow -2 \log c = \chi^2_{\alpha, 1} \\
 &\rightarrow \log c = -\frac{\chi^2_{\alpha, 1}}{2} \\
 &\rightarrow c = e^{-\chi^2_{\alpha, 1}/2}
 \end{aligned}$$

So, choose  $c$  in this way for *large*  $n$

- Wilke's Thm is for large sample size  $n$ . If  $n$  is small, the above cannot be used to choose  $c$  for the rejection region

- When  $n$  is small, we need to know the distribution of  $\Lambda$  under  $H_0$ ; this is called the **null distribution** of  $\Lambda$
- Ex: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$  where  $\theta > 0$ . Consider the hypothesis test  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$ . Then, the likelihood function for  $\theta$  is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\ &= \prod_{i=1}^n \frac{1}{\theta} e^{-X_i/\theta} \\ &= \theta^{-n} e^{-\frac{1}{\theta} \sum_{i=1}^n X_i} \end{aligned}$$

we have derived the MLE for  $\theta$  as  $\hat{\theta} = \bar{X}$  earlier. So, we now have

$$\begin{aligned} \Lambda &= \frac{L(\theta_0)}{L(\hat{\theta})} \\ &= \frac{\theta_0^{-n} e^{-\frac{1}{\theta_0} \sum_{i=1}^n X_i}}{\bar{X}^{-n} e^{-\frac{1}{\bar{X}} \sum_{i=1}^n X_i}} \\ &= \left( \frac{\bar{X}}{\theta_0} \right)^n e^{-\frac{n}{\theta_0} \bar{X} + n} \end{aligned}$$

We need to choose  $c$  such that  $\alpha = P(\text{Type I error}) = P(\text{reject } H_0 | H_0) = P(\Lambda \leq c | \theta = \theta_0)$ :

$$\begin{aligned} \Lambda \leq c &\iff \left( \frac{\bar{X}}{\theta_0} \right)^n e^{-\frac{n}{\theta_0} \bar{X} + n} \leq c \\ &\iff \left( \frac{\bar{X}}{\theta_0} \right)^n e^{-n \frac{\bar{X}}{\theta_0}} \equiv g\left(\frac{\bar{X}}{\theta_0}\right) \leq c_1 = ce^{-n} \end{aligned}$$

where  $g(t) = t^n e^{-nt}$ . As the log function is strictly increasing  $h(t) = \log g(t)$  attains its maximum value at the same point as  $g(t)$ , so we can solve  $g(t) \leq c_1$ :

$$\begin{aligned} g(t) &\leq c_1 \\ h(t) &= \log g(t) = n \log t - nt \\ \rightarrow h'(t) &= \frac{n}{t} - n \\ &= \begin{cases} > 0 & 0 < t < 1 \\ < 0 & t > 1 \end{cases} \end{aligned}$$

so  $h(t)$  is increasing on  $(0, 1)$  and decreasing on  $(1, \infty)$  (we note  $t = \left(\frac{\bar{X}}{\theta_0}\right)$  is always positive). Thus  $h(t)$  attains its maximum value at  $t = 1$  so  $g(t)$  also attains its maximum value at  $t = 1$ . Since  $g(0) = 0$ ,  $g(1) = e^{-n}$ , and  $g(\infty) = 0$  because

$$\begin{aligned} \lim_{t \rightarrow \infty} g(t) &= \lim_{t \rightarrow \infty} t^n e^{-nt} \\ &= \lim_{t \rightarrow \infty} \left( \frac{t}{e^n} \right)^n \\ &= \left( \lim_{t \rightarrow \infty} \frac{t}{e^n} \right)^n \end{aligned}$$

By L'Hopital's Rule, the above limit becomes  $\lim_{t \rightarrow \infty} \left( \frac{1}{e^n} \right)^n = 0$ . Now, note  $c_1 \leq ce^{-n} < e^{-n}$  because  $0 < c < 1$ . There exist  $0 < a < 1 < b$  such that  $g(t) < 1$  for  $t < a$  or  $t > b$ . Thus, note that

$$\Lambda \leq c \iff g\left(\frac{\bar{X}}{\theta_0}\right) \leq c_1 \iff \frac{\bar{X}}{\theta_0} \leq a \text{ or } \frac{\bar{X}}{\theta_0} \geq b \text{ which implies}$$

$P(\Lambda \leq c | H_0) = P\left(\frac{\bar{X}}{\theta_0} \leq a \text{ or } \frac{\bar{X}}{\theta_0} \geq b | \theta = \theta_0\right)$ . Consider  $Y_n = \frac{2}{\theta_0} \sum_{i=1}^n X_i$ ; its mgf  $M_{Y_n}(t) = E(e^{tY_n})$ . From homework 5, this is equal to



$$\begin{aligned}
E\left(e^{\frac{2t}{\theta_0} \sum_{i=1}^n X_i}\right) &= \prod_{i=1}^n E\left(e^{\frac{2t}{\theta_0} X_i}\right) \\
&= \prod_{i=1}^n M_{X_i}\left(\frac{2t}{\theta_0}\right) \\
&= \left(M_X\left(\frac{2t}{\theta_0}\right)\right)^n \\
&= \frac{1}{1 - \theta_0 \left(\frac{2t}{\theta_0}\right)^n}, \frac{2t}{\theta_0} < \frac{1}{\theta_0} \\
&= \frac{1}{(1 - 2t)^n}
\end{aligned}$$

Thus, under the null hypothesis  $H_0 : \theta = \theta_0$ , we know  $Y_n = \frac{2}{\theta_0} \sum_{i=1}^n X_i = \frac{2n}{\theta_0} \bar{X} \sim \chi_{(2n)}^2$ . Thus,

$$\begin{aligned}
\alpha &= P\left(\frac{\bar{X}}{\theta_0} \leq a \text{ or } \frac{\bar{X}}{\theta_0} \geq b \mid \theta = \theta_0\right) \\
&= P\left(Y_n = \frac{2n}{\theta_0} \bar{X} \leq 2na \text{ or } Y_n \geq 2nb \mid \theta = \theta_0\right) \\
&= P(\chi_{(2n)}^2 \leq 2na) + P(\chi_{(2n)}^2 \geq 2nb) \\
&\rightarrow 2na = \chi_{1-\frac{\alpha}{2}, 2n}^2, 2nb = \chi_{\frac{\alpha}{2}}^2 \\
&\rightarrow a = \frac{\chi_{1-\frac{\alpha}{2}}^2}{2n}, b = \frac{\chi_{\frac{\alpha}{2}}^2}{2n}
\end{aligned}$$

Hence for given  $\alpha \in (0, 1)$ ,  $\Lambda \leq c \iff \frac{\bar{X}}{\theta_0} \leq a \text{ or } \frac{\bar{X}}{\theta_0} \geq b$  where  $a = \frac{\chi_{1-\frac{\alpha}{2}}^2}{2n}, b = \frac{\chi_{\frac{\alpha}{2}}^2}{2n}$ , which gives  
 $\alpha = P(\Lambda \leq c \mid H_0) = P\left(\frac{\bar{X}}{\theta_0} \leq a \mid \theta = \theta_0\right) + P\left(\frac{\bar{X}}{\theta_0} \geq b \mid \theta = \theta_0\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$

## Sufficiency

- A point estimator  $Y_n = u(X_1, \dots, X_n)$  is called a **minimum variance unbiased estimator** (MVUE) of parameter  $\theta$  if  $E(Y_n) = \theta$  and  $\text{Var}(Y_n)$  is the smallest among all unbiased estimators for  $\theta$ 
  - The MVUE is sometimes called the **best** estimator
- What is MVUE's relation with efficient estimator?
  - Recall  $Y_n$  is an efficient estimator of  $\theta$  if 1)  $E(Y_n) = \theta$  ( $Y_n$  is an unbiased estimator of  $\theta$ ) and 2)  $\text{Var}(Y_n)$  is equal to the R-C lower bound
  - Thus, the difference between an efficient estimator and the MVUE is in their variance;  $\text{Var}(Y_n)$  for the MVUE  $Y_n$  may not be equal to the R-C lower bound
  - Note:
    - Sometimes the variance of the MVUE is higher than the R-C lower bound
    - If an estimator is efficient, then it is the MVUE
    - If an estimator is the MVUE, it may not be efficient
- How do we find the MVUE?
  - We will use **sufficient statistics**
- Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(x; \theta)$ . If  $Y = U(X_1, \dots, X_n)$  is a statistic with p.d.f.  $g(Y; \theta)$ , then  $Y$  is a **sufficient statistic** of  $\theta$  if and only if  $H(X_1, \dots, X_n) = \frac{L(\theta)}{g(U(X_1, \dots, X_n); \theta)}$  does not depend on  $\theta \in \Omega$  for every fixed value of  $Y = U(X_1, \dots, X_n)$ , where  $L(\theta) = \prod_{i=1}^n f(X_i; \theta)$ 
  - The sufficient statistic is not unique
  - Thm (Factorization Theorem): Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(X_i; \theta), \theta \in \Omega$ . Then  $Y = U(X_1, \dots, X_n)$  is a sufficient statistic of  $\theta$  if and only if we can find two nonnegative functions  $k_1, k_2$  such that  $L(\theta) = \prod_{i=1}^n f(X_i; \theta) = k_1(U(X_1, \dots, X_n); \theta) k_2(X_1, \dots, X_n) = k_1(Y; \theta) k_2(X_1, \dots, X_n)$
  - Examples

- Ex 1: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ . Then

$$f(X; \theta) = P_\theta(X = x) = \begin{cases} \theta^x (1 - \theta)^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

We know  $Y = X_1 + X_2 + \dots + X_n \sim B(n, \theta)$ . The p.d.f. of  $Y$  is given by

$$g(Y; \theta) = P(Y = y) = \begin{cases} \binom{n}{y} \theta^y (1 - \theta)^{n-y} & y = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

Consider the conditional probability of  $X_1, \dots, X_n$  given  $Y = y$ , e.g.

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n | Y = y) &= \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(Y = y)} \\ &\stackrel{\text{if } X_1 + \dots + X_n = Y}{=} \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(Y = y)} \\ &\stackrel{\text{independence of } X_i}{=} \frac{\prod_{i=1}^n P(X_i = x_i)}{g(Y; \theta)} \\ &= \frac{\prod_{i=1}^n \theta^{X_i} (1 - \theta)^{1-X_i}}{\binom{n}{y} \theta^y (1 - \theta)^{n-y}} \\ &= \frac{\theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}}{\binom{n}{y} \theta^y (1 - \theta)^{n-y}} \\ &= \frac{\theta^y (1 - \theta)^{n-y}}{\binom{n}{y} \theta^y (1 - \theta)^{n-y}} \\ &= \frac{1}{\binom{n}{y}} \end{aligned}$$

As  $\frac{1}{\binom{n}{y}}$  does not depend on  $\theta$ ,  $Y$  is a sufficient statistic of  $\theta$

- Ex 2: Let  $X_1, \dots, X_n$  be a random sample from the distribution with p.d.f.  $f(X; \theta) = e^{-(x-\theta)} I\{X > 0\}$  and let  $Y = \min(X_1, \dots, X_n)$ . Then the d.f. of  $Y$  is given by

$$\begin{aligned} G(Y; \theta) &= P(Y \leq y) \\ &= P(\min(X_1, \dots, X_n) \leq y) \\ &= 1 - P(\min(X_1, \dots, X_n) > y) \\ &= 1 - P(X_1 > y, \dots, X_n > y) \\ &\stackrel{\text{independence}}{=} 1 - \prod_{i=1}^n P(X_i > y) \\ &= 1 - \prod_{i=1}^n (1 - P(X_i \leq y)) \\ &\stackrel{\text{i.i.d.}}{=} 1 - (1 - F_X(y))^n \end{aligned}$$

Note

$$F(X; \theta) = \begin{cases} 1 - e^{-(x-\theta)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

As  $G(Y; \theta) = 1 - (1 - F_X(y))^n = 1 - (1 - (1 - e^{-(y-\theta)}))^n$ , we have

$$g(Y; \theta) = G'(Y; \theta) = \begin{cases} ne^{-n(y-\theta)} & y > \theta \\ 0 & y \leq \theta \end{cases}$$

so

$$\begin{aligned}\frac{\prod_{i=1}^n f(X_i; \theta)}{g(y; \theta)} &= \frac{\prod_{i=1}^n e^{-(x_i - \theta)} I\{X_i > \theta\}}{n e^{-n(y - \theta)} I\{y > \theta\}} \\ &= \frac{e^{\sum_{i=1}^n X_i + n\theta} I\{y > \theta\}}{n e^{-ny + n\theta} I\{y > \theta\}} \\ &= \frac{e^{-\sum_{i=1}^n X_i}}{n e^{-ny}}\end{aligned}$$

As this quantity does not depend on  $\theta$ , we have  $Y$  is a sufficient statistic of  $\theta$

- Ex 3: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$  where  $\sigma^2 > 0$  is known and  $\theta \in \mathbb{R}$ . Find a sufficient statistic for  $\theta$ 
  - Note that

$$\begin{aligned}\prod_{i=1}^n f(X_i; \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i - \theta)^2 / 2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \theta)^2\right)\end{aligned}$$

Recall

$$\sum_{i=1}^n (X_i - \theta)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \theta)^2$$

The proof of this is as follows:

$$\begin{aligned}\sum_{i=1}^n (X_i - \theta)^2 &= \sum_{i=1}^n ((X_i - \bar{X}) + (\bar{X} - \theta))^2 \\ &= \sum_{i=1}^n \left[ (X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \theta) + (\bar{X} - \theta)^2 \right] \\ &= \sum_{i=1}^n (X_i - \bar{X}) + 2(\bar{X} - \theta) \sum_{i=1}^n (X_i - \bar{X}) + n(\bar{X} - \theta)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X}) + n(\bar{X} - \theta)^2\end{aligned}$$

where we note  $\sum_{i=1}^n (X_i - \bar{X}) = 0$  by property of the sample mean  $\bar{X}$ . Using this identity, we have

$$\begin{aligned}&(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \theta)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 - n(\bar{X} - \theta)^2\right) \\ &= \exp\left(-\frac{n}{2\sigma^2} (\bar{X} - \theta)^2\right) (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= k_1(\bar{X}; \theta) k_2(X_1, \dots, X_n)\end{aligned}$$

where  $k_1(y; \theta) = \exp\left(-\frac{n(\bar{X} - \theta)^2}{2\sigma^2}\right)$  and  $k_2(x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\right)$ . Thus, by the

Factorization Thm, we know  $Y = \bar{X}$  is a sufficient estimator for  $\theta$

- Ex 4: Let  $X_1, \dots, X_n$  be a random sample from a distribution with p.d.f.

$$f(x; \theta) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find a sufficient estimator for  $\theta$

- We have

$$\begin{aligned}
 L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\
 &= \prod_{i=1}^n \theta X_i^{\theta-1} I\{0 < X_i < 1\} \\
 &= \theta^n \left( \prod_{i=1}^n X_i \right)^{\theta-1} \prod_{i=1}^n I\{0 < X_i < 1\}
 \end{aligned}$$

Let  $k_1(Y; \theta) = \theta^n \left( \prod_{i=1}^n X_i \right)^{\theta-1}$  and  $k_2(X_1, \dots, X_n) = \prod_{i=1}^n I\{0 < X_i < 1\}$  where  $Y = \prod_{i=1}^n X_i$  and  $k_1(y; \theta) = \theta^n y^{\theta-1}$  and  $k_2(X_1, \dots, X_n) = \prod_{i=1}^n I\{0 < X_i < 1\}$ . From the Factorization Thm, we know  $Y = \prod_{i=1}^n X_i$  is a sufficient statistic for  $\theta$ . Remark: going back to example 2, where  $f(x; \theta) = e^{-(x-\theta)} I\{X > \theta\}$ , we have

$$\begin{aligned}
 L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\
 &= \prod_{i=1}^n e^{-X_i+\theta} I\{X_i > \theta\} \\
 &= e^{n\theta - \sum_{i=1}^n X_i} \prod_{i=1}^n I\{X_i > \theta\}
 \end{aligned}$$

Let  $Y = \min(X_1, \dots, X_n)$ . Then the above expression becomes  $e^{n\theta - \sum_{i=1}^n X_i} I\{Y > \theta\} = e^{n\theta} e^{-\sum_{i=1}^n X_i} I\{Y > \theta\}$ . Then, we can let  $k_1(Y; \theta) = e^{n\theta} I\{Y > \theta\}$  and  $k_2(X_1, \dots, X_n) = e^{-\sum_{i=1}^n X_i}$ . Thus, by the Factorization Thm, we know  $Y$  is a sufficient statistic for  $\theta$

- Ex 5: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(0, \theta)$ . Do the following:
  - (a) Find a sufficient statistic  $Y_i$  for  $\theta$
  - We have

$$\begin{aligned}
 L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta}} e^{-X_i^2/2\theta} \\
 &= (2\pi\theta)^{-n/2} e^{-\frac{1}{2\theta} \sum_{i=1}^n X_i^2} \\
 &= k_1(Y; \theta) k_2(X_1, \dots, X_n)
 \end{aligned}$$

where  $Y_1 = \sum_{i=1}^n X_i^2$  and  $k_1(y; \theta) = (2\pi\theta)^{-n/2} e^{-\frac{y}{2\theta}}$  and  $k_2(X_1, \dots, X_n) = 1$ . By the Factorization Thm, we know  $Y_1 = \sum_{i=1}^n X_i^2$  is a sufficient statistic of  $\theta$

- (b) Show the MLE  $\hat{\theta}$  is a function of  $Y_1$
- We have

$$\begin{aligned}
 L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\
 &= (2\pi\theta)^{-n/2} e^{-\frac{1}{2\theta} \sum_{i=1}^n X_i^2} \\
 \implies l(\theta) &= \log L(\theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \theta - \frac{1}{2\theta} Y_1 \\
 \implies l'(\theta) &= -\frac{n}{2\theta} + \frac{Y_1}{2\theta^2} \\
 &= \frac{n}{2\theta^2} \left( \frac{Y_1}{n} - \theta \right) \\
 &= \begin{cases} > 0 & 0 < \theta < \frac{Y_1}{n} \\ < 0 & \theta > \frac{Y_1}{n} \end{cases}
 \end{aligned}$$

This means  $l(\theta)$  is increasing on the interval  $\left(0, \frac{Y_1}{n}\right)$  and decreasing on the interval  $\left(\frac{Y_1}{n}, \infty\right)$ . Thus,

$l(\theta)$  attains its maximum value at  $\theta = \frac{Y_1}{n}$  and the MLE of  $\theta$  is given by  $\hat{\theta} = \frac{Y_1}{n} = \frac{1}{n} \sum_{i=1}^n X_i^2$  which is a function of  $Y_1$

- (c) Find an unbiased estimator  $Y_2$  for  $\theta$ 
  - We first check if the MLE is an unbiased estimator:

$$\begin{aligned} E(\hat{\theta}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i^2) \\ &= \frac{1}{n} \sum_{i=1}^n \theta = \theta \end{aligned}$$

So the MLE  $\hat{\theta}$  is an unbiased estimator of  $\theta$ . We also know the sample variance  $Y_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$  is an unbiased estimator of  $\theta$ , but  $Y_2$  is not a function of  $Y_1$

- (d) Find  $\varphi(y_1) = E(Y_2|y_1)$  and compute  $\text{Var}(\varphi(Y_1))$
- To find  $\phi(y_1) = E(Y_2|Y_1)$ , we have

$$\begin{aligned} Y_2 &= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1} \\ &= \frac{Y_1 - n\bar{X}^2}{n-1} \end{aligned}$$

From the Student's Thm, we have  $Y_2 \perp \bar{X}$  implies  $Y_2 \perp \bar{X}^2$  and  $\frac{(n-1)Y_2}{\theta} \sim \chi_{(n-1)}^2$ . We also have

$\bar{X} \sim \mathcal{N}(0, \theta/n)$  implies  $\frac{\bar{X}}{\sqrt{\theta/n}} \sim \mathcal{N}(0, 1)$ , so  $\frac{\bar{X}^2}{\theta/n} \sim \chi_{(n-1)}^2$ . Thus,  $U = \frac{(n-1)Y_2}{\theta} \sim \chi_{(n-1)}^2$  and

$W = \frac{\bar{X}^2}{\theta/n} = \frac{Y_1 - (n-1)Y_2}{\theta} \sim \chi_{(n-1)}^2$ . This means the p.d.f. of  $(U, W)$  satisfies  $f(U, W) = f_U(u)f_W(w)$ . From

this, we can obtain the joint p.d.f. of  $g(y_1, y_2)$  of  $(Y_1, Y_2)$  using calculus. Finally, we can compute the conditional p.d.f.  $g(y_2|y_1)$ , which we can use to obtain  $\varphi(y_1) = \frac{y_1}{n}$  and  $\varphi(Y_1) = \frac{Y_1}{n} = \hat{\theta}$ . This gives

$E(\varphi(y_1)) = \theta$ . As  $\frac{(n-1)Y_2}{\theta} \sim \chi_{(n-1)}^2$ ,  $\text{Var}\left(\frac{(n-1)Y_2}{\theta}\right) = 2(n-1)$ , so  $\frac{(n-1)^2}{\theta^2} \text{Var}(Y_2) = 2(n-1)$ . So,

$\text{Var}(Y_2) = \frac{2\theta^2}{n-1}$  and

$\text{Var}(\varphi(y_1)) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n \sum_{i=1}^n X_i^2\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i^2) = \frac{1}{n^2} \sum_{i=1}^n 2\theta^2 = \frac{n(2\theta^2)}{n^2} = \frac{2\theta^2}{n} < \text{Var}(Y_2)$  This is because  $X_i \sim \mathcal{N}(0, \theta)$  implies  $\frac{X_i}{\sqrt{\theta}} \sim \mathcal{N}(0, 1)$ , which in turn implies  $\frac{X_i^2}{\theta} \sim \chi_{(1)}^2$  and  $\text{Var}\left(\frac{X_i^2}{\theta}\right) = 2$

## Properties of a Sufficient Statistic

- Thm: Let  $X_1, \dots, X_n$  be a random sample from a distribution with p.d.f  $f(x; \theta), \theta \in \Omega$ . If a sufficient statistic  $Y = u(X_1, \dots, X_n)$  for  $\theta$  exists and  $\hat{\theta}$  is the unique MLE of  $\theta$ , then  $\hat{\theta}$  is a function of  $Y$
- Thm: Let  $X$  and  $Y$  be r.v.s and  $E(Y) = \theta$ . Denote  $E(Y|X = x) = \varphi(x)$ . Then,

$$\begin{cases} E(\varphi(x)) = \theta \\ \text{Var}(\varphi(x)) \leq \text{Var}(Y) \end{cases}$$

- Remark: this theorem gives us a way to find a better unbiased estimator based on an existing unbiased estimator
- How do we choose  $X$  in practice?

- Thm (Rao-Blackwell Theorem): Let  $X_1, \dots, X_n$  be a random sample from a distribution with p.d.f  $f(x; \theta), \theta \in \Omega$ . If  $Y_1 = U_1(X_1, \dots, X_n)$  is a sufficient statistic for  $\theta$  and  $Y_2 = U_2(X_1, \dots, X_n)$  is an unbiased estimator of  $\theta$  but not a function of  $Y_1$ , then the following hold:

- (i)  $E(Y_2|Y_1 = y_1) = \varphi(y_1)$  defines a statistic  $\varphi(Y_1)$
- (ii)  $\varphi(y_1)$  is a function of a sufficient statistic of  $\theta$  and  $E(\varphi(y_1)) = \theta$
- (iii)  $\text{Var}(\varphi(y_1)) < \text{Var}(Y_2)$

# Completeness and Uniqueness

- We start with an example to introduce these concepts

- Ex 6: In problem 7.2.2, we have  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$  and know that  $Y_1 = \sum_{i=1}^n X_i$  is a **sufficient statistic** for  $\theta$ . The p.d.f. of  $X_i$  is

$$f(X_i; \theta) = \begin{cases} \frac{\theta^x e^{-\theta}}{x_i!} & x = 0, 1, \dots, \\ 0 & \text{otherwise} \end{cases}$$

The m.g.f. of  $X_i$  is

$$\begin{aligned} M_{X_i}(t) &= E(e^{tX}) \\ &= e^{\theta(e^t - 1)} \\ \Rightarrow M_{Y_1}(t) &= E\left(e^{t \sum_{i=1}^n X_i}\right) \\ &\stackrel{\text{independence}}{=} \prod_{i=1}^n E(e^{tX_i}) \\ &= \prod_{i=1}^n m_{X_i}(t) \\ &\stackrel{\text{i.i.d.}}{=} [M_X(t)]^n \\ &= e^{n\theta(e^t - 1)}, t \in \mathbb{R} \end{aligned}$$

This implies that  $Y_1 \sim \text{Poisson}(n\theta)$ . The p.d.f. of  $Y_1$  is then given by

$$g(y; \theta) = \begin{cases} \frac{(n\theta)^y e^{-n\theta}}{y!} & y = 0, 1, \dots, \\ 0 & \text{otherwise} \end{cases}$$

Consider the family of p.d.f.s  $\{g(y; \theta) | \theta > 0\}$ . Suppose  $u(Y)$  is a function of  $Y$  such that  $E(u(Y)) = 0$  for all  $\theta > 0$ . We will show  $u(0) = u(1) = \dots = 0$ , which means  $\{g(y; \theta) | \theta > 0\}$  is a **complete family** and  $Y$  is a **complete sufficient statistic** for  $\theta$ . We prove this as follows:

$$\begin{aligned} E(u(Y)) &= \sum_{y=0}^{\infty} u(y)g(y; \theta) \\ &= \sum_{k=0}^{\infty} u(k) \frac{(n\theta)^k e^{-n\theta}}{k!} \\ &= e^{-n\theta} \left( u(0) + u(1) \frac{n\theta}{1!} + u(2) \frac{(n\theta)^2}{2!} + \dots \right) \end{aligned}$$

from  $E(u(Y)) = 0$ , we know  $0 = e^{-n\theta} \left( u(0) + u(1) \frac{n\theta}{1!} + u(2) \frac{(n\theta)^2}{2!} + \dots \right)$  for all  $\theta > 0$ . Now,

note that  $\sum_{k=0}^{\infty} a_k x^k = 0$  for all  $x \in \mathbb{R}$  implies  $a_k = 0$  for  $k = 0, 1, 2, \dots$ . From this fact, we know

$$u(0) = nu(1) = \frac{n^2 u(2)}{2} = \dots = \frac{n^k u(k)}{k!} = \dots = 0, \text{ which implies } u(0) = u(1) = \dots = 0$$

so  $u(x) = 0$  on the support of  $Y$  and  $Y$  is a complete sufficient statistic for  $\theta$

- Let r.v.  $Z$  have a p.d.f.  $h(\xi; \theta)$ ,  $\theta \in \Omega$ . Then, the p.d.f family of  $\{h(\xi; \theta); \theta \in \Omega\}$  is called a **complete family** if  $[E(u(Z)) = 0 \forall \theta \in \Omega] \implies [u(\xi) = 0, \xi \text{ in the support of } h(\xi; \theta)]$  where the support of  $h(\xi; \theta) = \{\xi; h(\xi; \theta) \neq 0\} = \mathcal{S}$ .
- Thm (Lehmann-Scheffe Theorem): Let  $X_1, \dots, X_n$  be a random sample with p.d.f.  $f(x; \theta)$ ,  $\theta \in \Omega$  and let  $Y = u(X_1, \dots, X_n)$  be a sufficient statistic for  $\theta$ . Assume the p.d.f. of  $Y$  is  $g(y; \theta)$  and family  $\{g(y; \theta); \theta \in \Omega\}$  is complete. Then, if a function of  $Y$ , say  $\varphi(Y)$  is an unbiased estimator of  $\theta$  this  $\varphi(Y)$  must be the *unique* MVUE of  $\theta$ 
  - Ideas on how to find MUVE for  $\theta$  if  $X_1, \dots, \stackrel{\text{i.i.d.}}{\sim} f(X_i; \theta)$ 
    - Step 1: Find a sufficient statistic  $Y$  for  $\theta$
    - Step 2: Show  $Y$  is a complete sufficient statistic
    - Step 3: For any unbiased estimator  $W$  for  $\theta$  compute  $\varphi(u) = E(W|Y = y)$
    - Step 4:  $\varphi(Y)$  is the MVUE for  $\theta$
- Laplace transform

- Def: The **Laplace transform** of  $u(t)$  is defined as

$$\mathcal{L}(s) = \mathcal{L}(u(\cdot)) = \int_0^{\infty} u(t)e^{-st} dt$$

The Laplace transform is unique because

$$\left\{ \begin{array}{l} \int_0^{\infty} u(t)e^{-st} dt = 0 \quad \forall s > 0 \\ \int_0^{\infty} 0 \cdot e^{-st} dt = 0 \quad \forall s > 0 \end{array} \right.$$

- Examples

- Ex 7: Let  $Z \sim \text{Exp}(\theta)$ . Then  $f(\xi; \theta) = \begin{cases} \frac{1}{\theta} e^{-\xi/\theta} & \xi > 0 \\ 0 & \text{otherwise} \end{cases}$ . Consider the p.d.f family  $\{f(\xi; \theta) | \theta > 0\}$ . Assume  $E(u(Z)) = 0$ . Show  $u(\xi) > 0$  for all  $\xi > 0$ .
  - Sol: First note

$$\begin{aligned} E(u(Z)) &= \int_{-\infty}^{\infty} u(\xi) f(\xi; \theta) d\xi \\ &= \int_0^{\infty} u(\xi) \frac{1}{\theta} e^{-\xi/\theta} d\xi \end{aligned}$$

This requires the Laplace transform. From

$$E(u(Z)) = \frac{1}{\theta} \int_0^{\infty} u(t) e^{-t/\theta} dt = \frac{1}{\theta} \mathcal{L}_{\frac{1}{\theta}}(u(\cdot)) = 0 \quad \forall \theta > 0$$

Thus, by the uniqueness property of Laplace transforms, we have  $u(t) = 0 \quad \forall t > 0$ , so the p.d.f. family  $\{f(\xi; \theta) | \theta > 0\}$  is complete.

- Ex 8: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} U(0, \theta)$ . Then,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(X_i; \theta) \\ &= \prod_{i=1}^n \frac{1}{\theta} I\{0 < X_i < \theta\} \\ &= \theta^{-n} \left[ \prod_{i=1}^n I\{0 < X_i\} \right] \left[ \prod_{i=1}^n I\{X_i < \theta\} \right] \end{aligned}$$

Let  $Y_{(1)} = \min(X_1, \dots, X_n)$  and  $Y_{(n)} = \max(X_1, \dots, X_n)$ . The likelihood function then becomes

$$\theta^{-n} I\{Y_{(1)} > 0\} I\{Y_{(n)} < \theta\}$$

where we can let  $k_1(y; \theta) = \theta^{-n} I\{y < \theta\}$  and  $k_2(X_1, \dots, X_n) = I\{Y_{(1)} > 0\}$ . From the Factorization Theorem, we know  $Y_{(1)} = \min(X_1, \dots, X_n)$  is a sufficient statistic for  $\theta$ . From the previous example, we know the p.d.f. of  $Y_{(1)}$  is given by

$$g(y; \theta) = \begin{cases} \frac{ny^{n-1}}{\theta^n} & 0 < y < \theta \\ 0 & \text{otherwise} \end{cases}$$

We want to show  $Y_1$  is a complete sufficient statistic for  $\theta$ . Assume  $E(u(Y_{(1)})) = 0 \quad \forall \theta > 0$ . We want to show

$u(y) = 0 \forall y > 0$ . We have

$$\begin{aligned}
 0 & \stackrel{\forall \theta > 0}{=} E(u(Y_{(1)})) \\
 &= \int_0^\infty u(y)g(y; \theta) dy \\
 &= \int_0^\theta u(y) \frac{ny^{n-1}}{\theta^n} dy \\
 &= \frac{n}{\theta^n} \int_0^\theta u(y)y^{n-1} dy \\
 &\iff \int_0^\theta u(y)y^{n-1} dy = 0 \forall \theta > 0
 \end{aligned}$$

By the Fundamental Theorem of Calculus, we have

$$\frac{d}{d\theta} \left( \int_0^\theta u(y)y^{n-1} dy \right) = \frac{d}{d\theta} 0 = 0 \quad \forall \theta > 0$$

## Exponential Families

- Consider a special family of p.d.f.s  $\{f(x; \theta) \mid \theta \in \Omega\}$  where  $\Omega = \{\theta \mid \gamma < \theta < \delta\}$  where  $\gamma, \delta$  are constants or  $\pm\infty$ , and

$$f(x; \theta) = \begin{cases} \exp\{p(\theta)k(x) + H(x) + g(\theta)\}, & x \in \mathbb{R} \\ 0 & \end{cases}$$

- A p.d.f. with the form above is said to be a member of the **exponential class** of p.d.f.s
- A **nontrivial function** is a function equivalent to a constant everywhere
- A p.d.f. is said to be a **regular case of the exponential class** if:
  - The support  $S$  does not depend on  $\theta$
  - $p(\theta)$  is a nontrivial continuous function of  $\theta \in \Omega$
  - If  $X$  is a continuous r.v.,  $k'(x) \neq 0$ ,  $H(x)$  is continuous in  $x \in S$ . If  $X$  is a discrete r.v.,  $k(x)$  is a nontrivial function of  $x \in S$
- Let  $X_1, \dots, X_n$  be a random sample from a distribution that has a p.d.f. which is a regular case of the exponential class of the continuous type. Then,

$$\begin{aligned}
 \prod_{i=1}^n f(X_i; \theta) &= \exp\{p(\theta) + \sum_{i=1}^n k(X_i) + \sum_{i=1}^n h(x_i) + ng(\theta)\} \\
 &= \exp\{p(\theta) \sum_{i=1}^n k(x_i) + ng(\theta)\} \exp\{\sum_{i=1}^n h(X_i)\}
 \end{aligned}$$

Let  $k_1(U(X_1, \dots, X_n); \theta) = \exp\{p(\theta) \sum_{i=1}^n k(X_i) + ng(\theta)\}$  and  $k_2(X_1, \dots, X_n) = \exp\{\sum_{i=1}^n H(X_i)\}$ . By the Factorization Thm, we know  $Y = \sum_{i=1}^n k(X_i)$  is a sufficient statistic for  $\theta$ . Thus, in example 9, we have  $Y = \sum_{i=1}^n X_i^2$  is a sufficient statistic for  $\theta$ .

- Thm: Let  $\{f(x; \theta) \mid \gamma < \theta < \delta\}$  be a regular case of the exponential class. If  $X_1, \dots, X_n$  is a random sample from  $f(x; \theta)$ , then  $Y = \sum_{i=1}^n k(X_i)$  is a complete sufficient statistic for  $\theta$  (note this holds for either discrete or continuous r.v.s, it is not specified here).
- Steps to find MVUE:
  - 1. Identify the regular class of the exponential family
  - 2. Find  $E(Y_1)$  with  $Y_1 = \sum_{i=1}^n k(X_i)$
  - 3. Find a function of  $Y_1$ , say  $\varphi(Y_1)$ , which is unbiased estimator for  $\theta$
  - Since  $Y_1$  is a complete sufficient statistic for  $\theta$ , Lehmann-Scheffe Thm implies  $\varphi(Y_1)$  is the unique MVUE for  $\theta$
- Examples



- Ex 9: Let  $X \sim \mathcal{N}(0, \theta)$ . Then  $f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta} = \exp \left\{ -\frac{x^2}{2\theta} - \log \sqrt{2\pi\theta} \right\} = \exp \left\{ -\left( \frac{1}{2\theta} x^2 - \log \sqrt{2\pi\theta} \right) \right\}$  with  $\Omega = (0, \infty)$ ,  $S = \mathbb{R}$ ,  $p(\theta) = -\frac{1}{2\theta}$ ,  $k(x) = x^2$  (so  $k'(x) = 2x \neq 0$ ),  $H(x) = 0$  (so it is continuous), and  $g(\theta) = \log \sqrt{2\pi\theta}$ . Hence, we know  $\{f(x; \theta) \mid \theta > 0\}$  is a regular case of the exponential class of the continuous type.
- Ex 10: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$  where  $\sigma^2$  is known. Then,  $S = \mathbb{R}$ ,  $\theta \in \mathbb{R}$ , so

$$\begin{aligned} f(x; \theta) &= \frac{1}{\sqrt{2\pi\theta}} \\ &= \dots = \exp \left\{ -\log \sqrt{2\pi\sigma^2} - \frac{x^2 - 2\theta + \theta^2}{2\sigma^2} \right\} \\ &= \exp \left\{ \frac{\theta}{\sigma^2} - \frac{x^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2} - \frac{\theta^2}{2\sigma^2} \right\} \end{aligned}$$

is a regular case of the exponential class because  $p(\theta) = \frac{\theta}{\sigma^2}$  is a nontrivial function and continuous in  $\theta \in \mathbb{R}$ , so  $k(x) = x$ ,  $k'(x) = 1 \neq 0$ ,  $H(x) = -\frac{x^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2}$  is continuous in  $x$ , and  $g(\theta) = -\frac{\theta^2}{2\sigma^2}$ . Hence,  $Y_1 = \sum_{i=1}^n X_i$  is a complete sufficient statistic for  $\theta$ . Finally, to find the MVUE for  $\theta$ , we must find a function  $Y_1 \varphi(Y_1)$ . Since  $E(Y_1) = n\theta$ , then  $\varphi(Y_1) = \frac{Y_1}{n} = \bar{X}$  is a function of  $Y_1$  and  $E(\varphi(Y_1)) = E(\bar{X}) = \theta$ , so  $\varphi(Y_1) = \bar{X}$  is the unique MVUE for  $\theta$ .

- Ex 11: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$ . Then,  $f(x; \theta) = \frac{\theta^x e^{-\theta}}{x!}$ ,  $x \in S = \{0, 1, \dots\}$ ,  $\theta \in \Omega = (0, \infty)$ . So,  $S$  does not depend on  $\theta$ . Rewrite  $f(x; \theta)$  as

$$f(x; \theta) = \exp\{(\log \theta)x_i - \log(X_i) - \theta\}$$

This is a regular case of the exponential class because  $p(\theta) = \log \theta$  (nontrivial and continuous),  $k(x) = x$  (nontrivial in  $x \in S$ ), and  $H(x) = -\log(X_i)$ ,  $g(\theta) = \theta$ . Thus,  $Y_1 = \sum_{i=1}^n X_i$  is a complete sufficient statistic for  $\theta$ . Since  $E(Y_1) = n\theta$ , then  $\varphi(Y_1) = \frac{Y_1}{n} = \bar{X}$  is a function of  $Y_1$  and  $E(\varphi(Y_1)) = E(\bar{X}) = \theta$ . Hence,  $\varphi(Y_1) = \bar{X}$  is the unique MVUE for  $\theta$ . Note: generally try to keep  $k(x)$  as simple as possible to make computing  $Y_1 = \sum_{i=1}^n k(X_i)$  as simple as possible

## Functions of a Parameter

- Let  $Y$  be a complete sufficient statistic for  $\theta$ , and let  $\tau = h(\theta)$  be a function of  $\theta$  where  $h(\cdot)$  is one-to-one. We are interested in finding the MVUE of  $\tau$ . From the Factorization Thm, we know  $\prod_{i=1}^n f(X_i; \theta) = k_1(Y; \theta)k_2(X_1, \dots, X_n)$ . As  $\tau = h(\theta)$  is one-to-one, we have the inverse of  $h(\cdot)$  exists and  $\theta = h^{-1}(\tau)$ , so the factorization becomes

$$k_1(Y; h^{-1}(\tau))k_2(X_1, \dots, X_n) = \tilde{k}_1(Y; \tau)k_2(X_1, \dots, X_n)$$

This means  $Y$  is a complete sufficient statistic for  $\tau$ . Thus, if  $\varphi(Y)$  is a function of  $Y$  (a complete statistic) and  $E(\varphi(Y)) = \tau$ , from Lehmann-Scheffe Thm, we know  $\varphi(Y)$  is the unique MVUE for  $\tau$

- Examples
  - Ex 12: Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ . Then,  $f(x; \theta) = \theta^x (1 - \theta)^{1-x}$ ,  $x = 0, 1$  and  $\theta \in \Omega = (0, 1)$  and  $S = \{0, 1\}$  does not depend on  $\theta$ . Also,

$$\begin{aligned} f(x; \theta) &= \exp\{\log(\theta)x + (1 - x)\log(1 - \theta)\} \\ &= \exp\{[\log \theta - \log(1 - \theta)]x + \log(1 - \theta)\} \end{aligned}$$

which is a regular class of the exponential class of distributions because

$$\begin{cases} p(\theta) = \log \theta - \log(1 - \theta) & \text{nontrivial and continuous in } \theta \\ k(x) = x & \text{nontrivial} \\ H(x) = 0 \\ q(\theta) = \log(1 - \theta) \end{cases}$$

Thus,  $Y = \sum_{i=1}^n X_i$  is a complete sufficient statistic for  $\theta$ . Since  $\bar{X} = \frac{Y}{n}$  satisfies  $E(\bar{X}) = \theta$ , we know  $\bar{X} = \frac{Y}{n}$  is the unique MVUE for  $\theta$ . Consider  $\tau = \frac{\theta(1-\theta)}{n} = \text{Var}(\bar{X})$ , which is a one-to-one function of  $\theta$  (must show this on a homework or exam). We want to find the MVUE for  $\tau$ . Consider  $W = \frac{\bar{X}(1-\bar{X})}{n}$ , which is a function of the complete

sufficient statistic  $Y$ . We now want to compute  $E(W)$ :

$$\begin{aligned}
 E(W) &= \frac{1}{n} E(\bar{X} - \bar{X}^2) \\
 &= \frac{1}{n} (\theta - E(\bar{X}^2)) \\
 &= \frac{1}{n} (\theta - (\text{Var}(\bar{X}) + E(\bar{X})^2)) \\
 &= \frac{1}{n} \left( \theta - \left( \frac{\theta(1-\theta)}{n} + \theta^2 \right) \right) \\
 &= \frac{1}{n} \left( \theta(1-\theta) - \frac{\theta(1-\theta)}{n} \right) \\
 &= \frac{1}{n} \left( 1 - \frac{1}{n} \right) \theta(1-\theta)
 \end{aligned}$$

Thus,  $\varphi(Y) = \frac{n}{n-1} W = \frac{\bar{X}(1-\bar{X})}{n-1} = \frac{\bar{Y}(1-\frac{\bar{Y}}{n})}{n-1}$  is a function of a complete sufficient statistic for  $\tau$ . Since  $E(\varphi(Y)) = \tau$ , by Lehman-Scheffe Thm, we know  $\varphi(Y) = \frac{\bar{X}(1-\bar{X})}{n-1}$  is the MVUE for  $\tau$

- Ex 13 (not on exam): Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ . Consider

$$\begin{aligned}
 \tau &= P(X \leq c) \\
 &= \int_{-\infty}^c \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} dx \\
 &= x - \theta \int_{-\infty}^{c-\theta} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
 &= \Phi(c - \theta)
 \end{aligned}$$

so  $\tau = \Phi(c - \theta) = P(X \leq c)$ . To find the MVUE for  $\tau$ , the method used for Example 12 does not work because the  $\Phi$  function does not have a closed form and is defined via an integral. Consider  $Y_2 = u(X_1) = I\{X_1 \leq c\}$ . Then  $E(Y_2) = P(X_1 \leq c) = \tau$ . We can show  $Y = \sum_{i=1}^n X_i$  is a complete statistic for  $\tau$ . From the Rao-Blackwell Thm, we can find  $\varphi(x) = E(Y_2 | \bar{X} = x)$ . Then,  $\varphi(\bar{X})$  is the MVUE for  $\tau$ . Finding  $\varphi(x)$  is tedious and involves using the mgf of the normal distribution. In a nutshell,  $\varphi(\bar{X}) = \varphi\left(\frac{Y}{n}\right) = \Phi\left(\frac{\sqrt{n}(c - \frac{Y}{n})}{\sqrt{n-1}}\right)$