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: $\mathcal{S} = \{x \mid 0 \leq x \leq 12\} = [0, 12]$, 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 $A_1, \ldots, A_n$ 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: $P(\{1\}) = \cdots = P(\{6\}) = \frac{1}{6}$
    - 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
      - $\mathcal{S} = \{s, f\}$ (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: $P(A|B) = \frac{P(A \cap B)}{B}$
    - 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 $\mathcal{S} = \{HH, HT, TH, TT\}$. 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 $P(B|A) = \frac{P(A \cap B)}{A} = \frac{1/4}{1/2} = \frac{1}{2}$, so $P(B|A) = P(A)$ 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 $A_1, A_2, \ldots$ is a **partition** of a sample space S if:
      i) $A_1, A_2, \ldots$ are pairwise disjoint
      ii) $\bigcup_{i=1}^{\infty} A_i = \mathcal{S}$
    - **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:

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

$$P(\text{exactly one red ball in 4 trials}) = \sum_{i=1}^{4} P(\text{exactly one red ball at position } i)$$
$$= 4 \cdot \frac{3}{10} \cdot \left(\frac{7}{10}\right)^3$$

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 $\{0, 1, \ldots\}$ 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 = $\text{Bin}(n, p)$
    - $f(x) = \binom{n}{x} p^x (1-p)^x$, $x = 0, 1, \ldots, n$
  - Geometric = $\text{Geom}(p)$
    - $f(x) = p(1-p)^x$, $x = 0, 1, \ldots$
  - Negative binomial
  - Hypergeometric
  - Poisson
    - $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, 2, \ldots$
- 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
    - $G(1, \mu) = f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & x > 0 \\ 0 & x \le 0 \end{cases}$
    - μ 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
  - $0 \le F(x) \le 1 \; \forall x \in \mathbb{R}$
  - $F(x)$ is nondecreasing and right-continuous
  - $F(x)$ is continuous (step) function if X is a continuous (discrete) r.v.
  - $P(a \le X \le b) = F(b) - F(a)$
  - $f(x) = \begin{cases} F(x) - F(x-) & X \text{ discrete} \\ F'(x) & X \text{ continuous} \end{cases}$
  - 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 $x \le 1$, $f(t) = 0$. Therefore, our integral becomes

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

So, our answer is

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

  - Ex (1.7.24)
    - We will first find $F_Y(y) = P(Y \le y) = P(\tan X \le y)$ 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 $P(\tan X \le y)$ is the same as finding $P(X \le \tan^{-1} y) = \int_{-\infty}^{\tan^{-1} y} f(x)\, dx$. 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 $\mathbb{E}\left(e^{tX}\right) = M_X(t)$ 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:
    - 1. $M_X'(0) = \mathbb{E}(X) = \mu_X$
    - 2. $M_X^{(k)} = \mathbb{E}(X^k)$ (this is the kth 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

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

    - $M_X'(t) = \frac{3}{8}e^t + \frac{6}{8}e^{2t} + \frac{3}{8}e^{3t} \to M_X'(0) = 3/2 = \mu_X$
  - 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 $\mathbb{E}[e^{tX}] = \sum_{x=1}^{\infty} f(x)e^{tx} = \frac{1}{2}e^t + \frac{1}{4}e^{2t} + \cdots$
      - This is a geometric series with common ratio $\frac{e^t}{2}$, now, we can use the geometric series formula, we have

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

      - 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 $M'_X(0) = 2$
    - $\mathrm{Var}(X) = M''_X(0) - 4$; $M''_X(t) = -\frac{2e^t(e^t+2)}{(e^t-2)^3}$, so $M''_X(0) = -\frac{6}{-1} = 6$ and $\mathrm{Var}(X) = 6 - 2 = 4$
    - Recall the Taylor/Maclaurin series of a function is given by

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

    - 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 μ and variance σ^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

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

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

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

    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

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

  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

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

  - Def: The **correlation coefficient** of X and Y is defined as $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$
    - Properties
      - $|\rho| \leq 1$
      - $\rho > 0$ implies X and Y are *positively correlated*, meaning for a sample $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ 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 $\mathrm{Cov}(X,Y) < 0$)
      - $|\rho| \approx 1$ implies that X and Y *approximately* have a linear relationship, meaning a sample $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ from (X,Y) gives a linear straight line
- Moment generating function of (X,Y)
  - $M_{X,Y}(s,t) = \mathbb{E}\left[e^{sX+tY}\right]$
  - $\mathbb{E}[X^kY^m] = \frac{\partial^{k+m}M(s,t)}{\partial s^k \partial t^m}\Big|_{s=0,t=0}$

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

  - 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
    - $\mathrm{Cov}(X,Y) = 0$ (**note**: $\mathrm{Cov}(X,Y) = 0$ *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

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

        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 $\mathrm{Var}(Y|X=1)$, we compute $\mathbb{E}[Y^2|X=1]$ and compute $\mathrm{Var}(Y|X=1) = \mathbb{E}[Y^2|X=1] - (\mathbb{E}[Y|X=1])^2 = 15/64$

    - 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

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

    - 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 $\mu_Y = \int_{-\infty}^{\infty}$
    - 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 σ_X^2 and σ_Y^2 and can use the results from above to find ρ_{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 $T=\sum_{i=1}^n a_i x_i = a_1x_1+\cdots+a_nx_n$ and $W=\sum_{j=1}^m b_jY_j = b_1Y_1+\cdots+b_mY_m$
    - Thm: $\mathbb{E}[T]=\sum_{i=1}^n a_i\mathbb{E}[X_i]$ if and $\mathbb{E}[W]=\sum_{j=1}^m b_j\mathbb{E}[Y_i]$
    - $\text{Cov}(T,W)=\sum_{i=1}^n\sum_{j=1}^m a_jb_j\text{Cov}(X_i,Y_j)$
    - $\text{Var}(T)=\text{Cov}(T,T)=\mathbb{E}(T^2)-(\mathbb{E}[T])=\sum_{i=1}^n a_i^2\text{Var}(X_i)+2\sum_{i<j}a_ia_j\text{Cov}(X_i,X_j)^2$
    - 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 $\mathrm{Cov}(X,Y) = 3/4$
  - 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 $\mathrm{Var}(X) = 2r$
  - Poisson distribution
    - If $X \sim \mathrm{Poisson}(\lambda)$, $\mathbb{E}(X) = \lambda = \mathrm{Var}(X)$
- Normal distributions
  - If r.v. X has a **normal distribution** with mean μ and variance σ^2, we denote $X \sim \mathcal{N}(\mu, \sigma^2)$
  - Properties
    - The m.g.f. of X is given by

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

  - Bivariate normal distribution
    - A random vector has a **bivariate normal distribution** $\mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)$ 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

$$M(t_1,t_2) = \mathbb{E}(e^{t_1X_1 + t_2X_2}) = e^{\mu_1t_1 + \mu_2t_2 + \frac{\sigma_1^2 + t_1^2 + 2\rho\sigma_1\sigma_2t_1t_2 + \sigma_2^2t_2^2}{2\left(1-\rho^2\right)}}$$

      - The above implies $M_X(t) = M(t,0) = \mathbb{E}(e^{tX+0\cdot y}) = e^{\mu_1t + \frac{\sigma_1^2}{2}t^2}$, 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 $\left.\frac{\partial^2 M}{\partial t_1 \partial t_2}\right|_{(0,0)} = \mathbb{E}(XY)$ for a bivariate normal distribution as follows:

$$\begin{aligned} \frac{\partial^2 M}{\partial t_1 \partial t_2} M(t_1,t_2) &= \frac{\partial}{\partial t_2}\left(\frac{\partial M}{\partial t_1}\right) M(t_1,t_2) \\ &= M(t_1,t_2)(\mu_1 + \sigma_1^2 t_1 + \rho\sigma_1\sigma_2t_2) + (\mu_2 + \sigma_2^2t_2 + \rho\sigma_2\sigma_2t_1) + M(t_1,t_2)\rho\sigma_2\sigma_2 \\ \to \mathbb{E}(XY) &= M(0,0)\mu_1 + \mu_2 + M(0,0)\rho\sigma_1\sigma_2 \\ &= \mu_1\mu_2 + \rho\sigma_1\sigma_2 \end{aligned}$$

    - Using the above result, we obtain $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mu_1\mu_2 + \rho\sigma_1\sigma_2 - \mu_1\mu_2 = \rho\sigma_1\sigma_2$
    - 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 $\mathrm{Cov}(X,Y) = 0$, but $\mathrm{Cov}(X,Y) = 0$ 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 $X_1, \ldots, X_n$ be mutually independent normal r.v.s. Then, for constants $a_1, \ldots, a_n$,

$$\begin{aligned} Y &= \sum_{i=1}^{n} a_i X_i \\ &= a_1X_1 + \cdots + a_nX_n \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right) \end{aligned}$$

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

  - Proof: From previous results, we have $\mu_Y = \sum_{i=1}^{N} a_i\mu_i$, $\sigma_Y^2 = \mathrm{Var}(Y) = \sum_{i=1}^{n} a_i^2\sigma_1^2$. 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

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

  Since $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, we know $M_{X_i}(t) = e^{\mu_it + \frac{\sigma_i^2}{2}}$. 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 $\mathcal{N}\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\right)$ is equal to the mgf of Y, implying $Y \sim \mathcal{N}\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\right)$, 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, $(\overline{X}, \overline{Y}) \sim N_2\left(\mu_1, \mu_2, \frac{\sigma_1^2}{n}, \frac{\sigma_2^2}{n}, \rho\right)$.
    - Proof: We will prove the mgf of both distributions is the same. To compute the mgf of $(\overline{X}, \overline{Y})$, we have

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

Doing some more algebra leads to the fact that the mgf of $(\overline{X}, \overline{Y})$ 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 $Z \sim \mathcal{N}(0, 1)$, $V \sim \chi^2_{(r)}$ (where $\chi^2_{(r)}$ 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

$$f(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\Gamma\left(\frac{3}{2}\right)}\frac{1}{\left(1+\frac{t^2}{r}\right)^{\frac{r+1}{2}}}$$

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

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

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

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

- 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

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

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

$$\left(\frac{X_1 - \mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \mu}{\sigma}\right) \sim \chi^2_{(n)}$$

      Since $\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{x}}{\sigma}\right)$, 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

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

- 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. $\mathrm{Exp}(\mu) = G(1, \mu)$, where $M_{G(1,\mu)}(t) = \frac{1}{1-\mu t}$, $t < \mu$
    - 2. $\chi^2_{(r)} = G\left(\frac{r}{2}, 2\right)$, $M(t) = \frac{1}{(1-2t)^{r/2}}$, \$t < \frac 12\$4

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 defns
  - Let $X_1, \ldots, X_n$ be n independent r.v.s each with d.f. $F(x)$. Then, $X_1, \ldots, X_n$ 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 $T = T(X_1, \ldots, X_n)$ be a function of a random sample $X_1, \ldots, X_n$. 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 $X_1, \ldots, X_n$
  - We say $\hat{\theta}$ is an **unbiased estimator** for θ 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 μ, 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 μ because $\hat{x}$ is close to μ 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 μ
      - Ideally, for a small $\delta > 0$,

        $$P(|\bar{x} - \mu| \le \delta) = 1 - \alpha$$

        where $0 < \alpha < 1$ is a small number
      - Ex: if we have $\delta = 0.01$ and $\alpha = 0.05$, the above equation becomes

        $$P(|\bar{x} - \mu| \le 0.01) = 0.95$$

        In words, the probability the estimation error does not exceed 0.01 is 0.95
      - Note the inequality $|\bar{x} - \mu|$ implies $-0.01 \le \bar{x} - \mu \le 0.01$, which can be rewritten as $\bar{x} - 0.01\mu \le \bar{x} + 0.01$ can be written as $P(\bar{x} - 0.01 \le \mu \le \bar{x} + 0.01)$
        - 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 μ
      - Question: for a given $\alpha > 0$, what is the value of $\delta > 0$ in $P(|\bar{x} - \mu| \le \delta) = 1 - \alpha$
        - 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,

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

          - 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 α is known
  - Let $X_1, \dots, X_n$ be a random sample from a distribution with parameter θ and let $a_n = a_n(X_1, \dots, X_n) \le b_n = b_n(X_1, \dots, X_n)$ 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 θ if $P(a_n \le \theta \le b_n) = 1 - \alpha$
  - Ex: 4.2.19
    - Let $Y = \frac{2}{\beta}\sum_{i=1}^{n} X_i$. Then,

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

```
For $t < \frac 12$, $\left(\frac{1}{\left[1-\beta\left(\frac{2t}
{\beta}\right)\right]^3}\right)^n = \frac{1}{(1-2t)^{3n}}$, so $M_Y(t) = \frac{1}{(1-
2t)^{3n}}$, where $t < \frac 12$, so $Y \sim \chi_{(6n)}$, which implies $1-\alpha =
P(\chi^2 _{1-\frac{\alpha}{2}, 6n} \leq \frac{2n}{\beta} \leq \chi^2 _{\frac{\alpha}{2},
6n})$
```

- Hypothesis testing
  - Ex: Let $\mu = [\text{average family income in MD}]$. 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 α and β are minimized (a small α is especially crucial due to the goal of hypothesis testing)
          - α 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 α 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 α is small, which is decided in advance (e.g. before you conduct a hypothesis test)
        - Choosing c
          - Recall if $X_1, \ldots, X_n$ is a random sample from a population, with mean μ and variance σ^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 α as follows:

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

        - 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 reject 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 α (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 μ; 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

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

    - This is denoted as $Y_n \xrightarrow{P} Y$ as $n \to \infty$
    - Note: such a definition means if $\hat{\theta}_n = Y_n = \hat{\theta}(X_1, \dots, X_n)$ is a point estimator for θ 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 $n \to \infty$
        - In such a case, we call the point estimator $\hat{\theta}_n$ a **consistent estimator** if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$
- Weak Law of Large Numbers (WLLN)
    - Thm (WLLN): If $X_1, \dots, X_n$ is a random sample from a distribution with mean μ and variance σ^2, then $\overline{x}_n \xrightarrow{P} \mu$ as $n \to \infty$
        - Proof: Recall $E(\bar{x}) = \mu$, $\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$. By Chebyshev's Inequality, we have $$\forall \epsilon > 0 \ P(|\bar{x}-\mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X})}{n\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}

$$As\$n \to \infty\$, theRHSgoestozero, sotheprobabilityalsoapproaches0, implying\$ \lim_{n\to\infty} P(|\bar{x}-\mu| \geq \epsilon) = 0\$, implying\$\bar{x}$$

        - Note: WLLN implies that $\bar{x}$ is a **consistent estimator** for μ
    - 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 $n \to \infty$
- Examples
    - Ex 1: Let $X_1, \dots, x_n$ be a random sample from a distribution with mean μ and variance σ^2. Assume $E(X_i^4) < \infty$. Show that the sample variance s^2 is a consistent estimator for σ^2
        - Sol: First, note it is relatively easy to prove that $\bar{x}$ is an unbiased estimator of μ 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

μ. By WLLN, we can prove $\bar{x} \xrightarrow{P} \mu$ as $n \to \infty$, so $\bar{x}$ is a consistent estimator of μ. To prove the sample variance is a consistent estimator of σ^2we 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 $\frac{n}{n-1} = \frac{1}{1-\frac{1}{n}} \to 1$ as $n \to \infty$. 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 σ^2. We now ask: is s_n^2 an unbiased estimator of σ^2?

- Ex: Let $X_1, \ldots, X_n$ be i.i.d. r.v.s with distribution $U(0, \theta)$. Let $Y_n = \max\{X_1, \ldots, X_n\} = X_{(n)}$. Show (a) Y_n is a biased estimator of θ and (b) Y_n is a consistent estimator of θ. (c) $W_n = \frac{n+1}{n}Y_n$ is an unbiased estimator of θ.
  - Sol: For part (a), we want to find the pdf of Y_n using the distribution function of each X_i. We have $F_{Y_n}(y) = P(Y_n \leq y) = P\{\max\{X_1, \ldots, X_n\} \leq y\}$. We note that by independence, $\prod_{i=1}^{n} P(X_i \leq y) = [F_X(y)]^n$, so

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

    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,

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

    Thus, Y_n is a biased estimator for θ. 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:

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

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

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

Therefore, $Y_n - E(Y_n) = Y_n - \frac{n}{n+1}\theta \xrightarrow{P} 0$ as $n \to \infty$. Also, $\frac{n}{n+1}\theta \to \theta$ as $n \to \infty$, 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 θ

- 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 $X_n, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. We know $\bar{x}$ is an unbiased estimator for μ and $\bar{x}$ is a consistent estimator of μ. 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 $\lim_{n\to\infty} F_{X_n}(x_0) = F_X(x_0)$ denoted by $X_n \xrightarrow{D} X$ as $n \to \infty$. 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 $\lim_{n\to\infty} M_{X_n}(t) = M(t)$ 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 $g(X_n) \to g(X)$ as $n \to \infty$
    - 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 $A_nX_n + B_n \xrightarrow{D}\to aX + b$ as $n \to \infty$
    - Useful limits:

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

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

  - 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 $\lim_{n\to\infty} F_{Y_n}(y) = \lim_{n\to\infty}\left(\frac{y}{\theta}\right)^n = 0$. 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 $\lim_{n\to\infty} F_{Y_n}(y) = \begin{cases} 0 & y \leq 0 \\ 0 & 0 < y < \theta \\ 1 & y \geq \theta \end{cases}$

Thus, $Y_n \xrightarrow{D} y$ as $n \to \infty$

- Examples
  - Ex: Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(0,1)$. Then, $Y_n = \overline{X}_n \sim \mathcal{N}(0, 1/n)$, 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

      $$\begin{aligned} \int_{-\infty}^{\sqrt{n}y} \sqrt{\frac{n}{2\pi}} e^{-u^2/2}\, du &= \int_{-\infty}^{\sqrt{n}y} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \\ &= \Phi\left(\sqrt{n}y\right) \end{aligned}$$

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

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

      So, we have

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

      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 $\lim_{n\to\infty} F_n(y) = F_Y(y)\ \forall y \neq 0$. Thus, $Y_n = \overline{X}_n \xrightarrow{D} Y$ as $n \to \infty$
  - 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 $n \to \infty$ 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. if 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

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

  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 $n \to \infty$, we have

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

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

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

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

- Ex: Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d.. Recall the Student's Theorem, which states $T_n = \frac{\bar{x} - \mu}{\sqrt{s_n^2/n}} \sim t_{(n-1)}$. 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 $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$. Note we previously showed $s^2 \xrightarrow{P} \sigma^2$ as $n \to \infty$. By the Continuity Theorem, $s_n \xrightarrow{P} \sigma$ since $G(x) = \sqrt{n}$ 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 $n \to \infty$, so $T_n \xrightarrow{D} \mathcal{N}(0,1)$, as desired.
- Ex: Let $Y_n \sim \mathrm{Bin}(n, p)$. Does Y_n converge in distribution? Assume $\mu = np$ remains the same since μ comes from a large population with a fixed mean.
  - Solution: Note $E(Y_n) = np$, $\mathrm{Var}(Y_n) = np(1-p)$. 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 $n \to \infty$, we have

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

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 $n \to \infty$. 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 ξ is between t_0 and t, the above expression becomes

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

So, because

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

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 $\left(1-\frac{2}{n}t^2+\frac{t^2}{n}+a_n\right)^{-n/2} = \left(1-\frac{t^2}{n}\right)^{-\frac{n}{2}}$. Now, we can take the limit as $n \to \infty$ to obtain $\lim_{n\to\infty} M_{Y_n}(t) = \lim_{n\to\infty}\left[\left(1+\frac{1}{-n/t^2}\right)^{-\frac{n}{t^2}}\right]^{\frac{t^2}{2}} = e^{t^2/2}$, which is the mgf of $\mathcal{N}(0,1)$. Thus, $Y_n \xrightarrow{D} \mathcal{N}(0,1)$ as $n \to \infty$

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

- 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 $\prod_{i=1}^{n} E\left(e^{\frac{t}{\sqrt{n}}W_i}\right) = \prod_{i=1}^{n} M_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right)$. 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 ξ_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 ξ_n is between 0 and $\frac{t}{\sigma\sqrt{n}}$. Since $\frac{t}{\sigma\sqrt{n}} \to 0$ and $\xi_n \to 0$ as $n \to \infty$, we have

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

  Hence, $\lim_{n\to\infty} M_{Y_n}(t) = e^{t^2/2}$, so $Y_n \xrightarrow{D} \mathcal{N}(0,1)$ as $n \to \infty$, as desired.
- Properties:
  - Let $Z \sim \mathcal{N}(0,1)$. By the Continuity Theorem,

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

Maximum Likelihood Methods

- Let $X_1, \dots, X_n \sim f(x;\theta_0)$, where $\theta_0 \in \Theta$ be i.i.d, where Θ is the parameter space for θ_0 and $f(x;\theta_0)$ is known except the value of θ_0. This is called a **parametric model**, since there is a parameter θ_0
  - How do we estimate θ_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 θ_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 ξ_i is between $x_i \pm \delta$, the above becomes

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

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

- 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 θ is given by $L(\theta) = \prod_{i=1}^{n} f(X_i;\theta_0)$

- Maximum likelihood estimators
  - The MLE $\hat{\theta}$ for θ_0 is given by $L(\hat{\theta}_n) = \max_{\theta\in\Theta} L(\theta)$
  - What is the meaning of the MLE $\hat{\theta}$ of θ_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 θ, then $\hat{\eta} = g(\hat{\theta})$ ist he MLE for for $\eta = g(\theta)$
  - Thm (Consistency Theorem): Under regularity (normal) conditions, $\hat{\theta} \to \theta$ as $n \to \infty$
- Examples
  - Ex: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ be i.i.d. Find the MLE for μ
    - Sol: The likelihood function for μ is given by

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

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

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

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

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

So, the likelihood function for μ 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 $l(\theta)$

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

    This means $l(\theta)$ is decreasing on the interval $[X_{(n)}, \infty)$, so $l(\theta)$ attains its maximum value at $\theta = X_{(n)}$, so the MLE for θ is given by $\hat{\theta} = X_{(n)} = \max(X_1, \dots, X_n)$

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

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

    Since the support of $f(x; \mu)$ does not depend on μ, 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

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

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

- Ex: Let $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} \text{Exp}(\mu)$. Find the MLE for μ 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 μ 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

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

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

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

- Ex: Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Find the MLE for μ and σ^2
  - Sol: Let $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^{\geq 0}$, where Θ 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 θ 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 μ 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 μ is given by $\mu = \bar{X}$. To find the MLE for σ^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 μ, $l(\theta)$ 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 μ and σ^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 σ^2 first and computed $\max_{\mu\in\mathbb{R}}\left(\max_{\sigma^2\in\mathbb{R}^+}\right)$, 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

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

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

Rao-Cramer Lower Bound and Efficiency

- Motivation: (a) What is a "good" estimator for θ? (b) What is the best point estimator for θ?
- We know that if a point estimator $\hat{\theta}$ for θ 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 θ by definition of expected value
  - If we have two point estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for θ, both are unbiased estimator, which one is better?
    - Since $E(\hat{\theta}_1) = \theta$ and $E(\hat{\theta}_2) = \theta$, we have

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

      - Since $E[(\hat{\theta}_i-\theta)^2]$ measures how much $\hat{\theta}_i$ *deviates from θ on average*, thus $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$, then $\hat{\theta}_1$ is a better estimator than $\hat{\theta}_2$ for θ
- In general, a point estimator $\hat{\theta}$ is considered to be good if $E(\hat{\theta}) = \theta$ and $\mathrm{Var}(\hat{\theta})$ is small
- Among all unbiased estimators for θ, 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 $I(\theta) = E\left[\left(\frac{\partial \log f(X;\theta)}{2\theta}\right)^2\right]$
    - Under some regularity conditions, we can show $I(\theta) = -E\left[\frac{\partial^2 \log f(X;\theta)}{2\theta^2}\right]$
- Rao-Cramer inequality
  - Thm (**Rao-Cramer inequality**): Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta)$, $\theta \in \Omega$. if $Y = u(X_1, \dots, X_n)$ is an unbiased estimator of θ, then $\sigma_Y^2 = \mathrm{Var}(Y) \geq \frac{1}{nI(\theta)}$ 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 $\frac{\text{R-C lower bound}}{\sigma_Y^2} = \frac{Y_{nI(\theta)}}{\mathrm{Var}(Y)}$
  - Def: Let Y be an unbiased estimator of θ. Then, Y is called an **efficient estimator** if and only if σ_Y^2 attains the R-C lower bound
  - Thm: If $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} = f(X_i;\theta_0)$ and $\hat{\theta}$ is the MLE of θ_0, then under regularity conditions, this $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$
- Examples
  - Ex: If $X_1, \dots, X_n$ is a random sample from a distribution with mean μ and variance σ^2. Let $\hat{\theta}_1 = x_1$, $\hat{\theta}_2 = \bar{x}$. Then $E(X_1) = \mu$, $E(\bar{X}) = \mu$. But

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

    Since $\mathrm{Var}(\hat{\theta}_2) < \mathrm{Var}(\hat{\theta}_1)$ for $n \geq 2$, we know $\bar{X} = \hat{\theta}_2$ is a better point estimator fro μ than $\hat{\theta}_1 = X_1$
  - Ex: Let $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$. In an earlier example, we shows 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 $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma} \sim \chi^2_{(n-1)}$. This implies

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

    and $\mathrm{Var}\left(\frac{n\hat{\sigma}}{\sigma^2}\right) = 2(n-1)$. 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 $E(\hat{\sigma}^2) \to \sigma^2$ as $n \to \infty$ by WLLN, so the MLE $\hat{\sigma}^2$ of σ^2 is not an unbiased estimator. Adjusting this, we have an unbiased estimator of σ^2: $s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ 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, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$, so we know $\mathrm{Var}\left(\frac{(n-1)s^2}{\sigma^2}\right) = 2(n-1)$, which implies

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

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

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

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

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

So,

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

and

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

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 $s^2 = \left(1 - \frac{1}{n}\right) \to 1$ as $n \to \infty$. 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 $n \to \infty$
  - (c) Sometimes, a biased estimator may be...