Random Variables and Expectation

Statistical Methods in Political Research I

Yuki Shiraito

University of Michigan

Fall 2018

Random Variables

- Data are numbers
- How do we link sample spaces and events to numbers?

- Implicitly we have used:
  - Dice roll: Each outcome has a number
  - Falling stick: Azimuth degrees from 0 to 360
  - Survey responses: Yes as 1, No as 0
  - Supreme court: Number of judges voting for the plaintiff

- Random variable: A *random variable* X is a function

$$X : \Omega \to \mathbb{R}$$

  such that for any real number $x \in \mathbb{R}$, $\{\omega \mid X(\omega) \leq x\} \in \mathcal{F}$

- Convention:
  1. Uppercase letter such as X stands for r.v.
  2. Lowercase letter such as x stands for a *realized value* of r.v.

Remarks on Random Variables

- Random variable is a *function*:
  - Takes an outcome in the sample space as an argument
  - Gives a single value assigned to each outcome
  - May give a common value for multiple outcomes

- Can consider different r.v.s for the same probability space:
  - Dice roll:
    - Numbers on the dice
    - -1 if 1 on the dice, 1 if 6 on the dice, 0 otherwise
    - 1 if an even number on the dice, 0 if an odd number on the dice
  - Survey responses:
    - 1 if "yes", 0 if "no" for each response
    - Number of respondents who answer "yes" (sum of the above)
    - Number of times a respondent answers "yes" (multiple responses)

- In applications, it is important to find a useful r.v.

Distribution

- *Distribution* of a random variable:
  - Let C be a subset of $\mathbb{R}$ such that $\{\omega \mid X(\omega) \in C\}$ is an event
    1. $\mathbb{P}(X \in C) \equiv \mathbb{P}(\{\omega \mid X(\omega) \in C\})$
    2. The *distribution* of X: The collection of $\mathbb{P}(X \in C)$ for all possible C
- Distribution of X can be considered as a probability measure:
  1. Sample space: $\mathbb{R}$
  2. Set of events: Set of all possible C
  3. Probability measure: $\mathbb{P}(X \in C)$
- Betting on even or odd numbers from a dice roll:
  - Sample space: 6 faces of a dice
  - Events: $\emptyset$, even, odd, all
  - Probability measures: $0, 1/2, 1/2, 1$
  - Random variable: $X(\omega) = 1$ if even, $X(\omega) = 0$ if odd
  - $\mathbb{P}(X \leq 0) \equiv \mathbb{P}(\text{odd}) = 1/2, \mathbb{P}(X > 0) \equiv \mathbb{P}(\text{even}) = 1/2$
  - $C \in \{\emptyset, \{r \in \mathbb{R} \mid r \leq 0\}, \{r \in \mathbb{R} \mid r > 0\}, \mathbb{R}\}$
- Will directly work with r.v.: Write X instead of $X(\omega)$
- Probability space is hidden behind r.v., but it's there

Cumulative Distribution Function

- Generally, there are a huge number of C
- Need a simple way to describe a distribution

- Cumulative distribution function: The *cumulative distribution function (c.d.f.)* of a r.v. X, denoted by F_X, is a function $F_X : \mathbb{R} \to [0, 1]$ such that
$$F_X(x) = \mathbb{P}(X \leq x)$$
- Remember the definition of r.v.: $\{\omega \mid X(\omega) \leq x\} \in \mathcal{F}$ for any $x \in \mathbb{R}$

- Example of c.d.f.:
  - Betting on even or odd numbers on a dice roll
  - Dice roll in the Nigeria Survey
  - Stick fall in the Nigeria Survey

- Uniqueness of c.d.f.: Let X have c.d.f. F and Y have c.d.f. G. If $F(x) = G(x)$ for all $x \in \mathbb{R}$, then $\mathbb{P}(X \in C) = \mathbb{P}(Y \in C)$ for all possible C.

Valid C.D.F.

- Properties of c.d.f.: A function $F : \mathbb{R} \to [0, 1]$ is a c.d.f. for some probability $\mathbb{P}$ if and only if F satisfies the following three conditions:
  1. F is non-decreasing: $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$
  2. F is normalized:
  $$\lim_{x \to -\infty} F(x) = 0$$
  $$\lim_{x \to \infty} F(x) = 1$$
  3. F is right-continuous: $F(x) = \lim_{y \downarrow x} F(y)$ for all x

  Proof. Samuel's section and future problem set.

- *Any* function satisfying the three conditions can be a c.d.f.
- Not necessarily well known

Discrete Random Variable

- Discrete r.v.: X is *discrete* if it takes only countably many values
- Probability function: For a discrete r.v. X, the *probability (mass) function (p.f.)* of X, denoted by $f_X : \mathbb{R} \to [0,1]$, is defined by $f_X(x) \equiv \mathbb{P}(X = x)$
- Support of X: $\{x \in \mathbb{R} \mid f_X(x) > 0\}$
- C.d.f. and p.f.:

$$F_X(x) = \sum_{\{y \mid f_X(y) > 0 \,\wedge\, y \leq x\}} f_X(y)$$

$$f_X(x) = \lim_{y \downarrow x} F_X(y) - \lim_{y \uparrow x} F_X(y)$$

- X is discrete $\Leftrightarrow$ c.d.f. of X is a step function
- Valid p.f.: The p.f. of X with its support $\{x_1, \ldots\}$ must satisfy the following two conditions:
  1. f is non-negative: $f_X(x) \geq 0$
  2. f sums to 1: $\sum_{i=1}^{\infty} f_X(x_i) = 1$

Bernoulli Distribution

- *Bernoulli trial*: Random realization of a "success" or a "failure"
  - Survey response to a yes/no question
  - Yea/nay vote by a legislator, judge, representative, ...
  - Any binary feature (e.g. democracy, below/above a threshold, etc.)
- X follows a Bernoulli distribution with support $\{0, 1\}$:

$$F_X(x) = \begin{cases} 0 & (x < 0) \\ 1 - p & (0 \leq x < 1) \\ 1 & (1 \leq x) \end{cases}$$

$$f_X(x) = p^{1\{x=1\}}(1-p)^{1\{x=0\}}$$

  - *Parameter* of the Bernoulli distribution: $p = \mathbb{P}(X = 1)$
  - $1\{\cdot\}$: Indicator function
  - Denoted by: $X \sim \text{Bern}(p)$
- C.d.f. and p.f. of known distributions:
  - Values of X and parameters
  - $F_X(x; \theta)$, $f_X(x; \theta)$

Binomial Distribution

- Sum of "successes" in n independent Bernoulli trials
  - Nigeria survey: Number of "yes" answers if asked multiple times
  - Conflict example in Pset 2: Number of battles a country wins
  - Shop owner: Number of customers who give 3+ stars on Yelp
- X follows a Binomial distribution with support $\{0, 1, \ldots\}$:

$$F_X(x; n, p) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

$$f_X(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

  - *Parameters* of the Binomial distribution: p and n
  - $\lfloor x \rfloor$: the greatest integer less than or equal to x
  - Denoted by: $X \sim \text{Binom}(n, p)$

- Story gives the p.f. of the Binomial distribution
- Binomial r.v. is the sum of Bernoulli r.v.
- Will revisit the transformation of r.v.s

Continuous Random Variable

- Probability density function: If there exists a function $f_X : \mathbb{R} \to \mathbb{R}^+$ such that for any $a, b \in \mathbb{R}$ with $a \leq b$,
$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$
and $\int_{-\infty}^{\infty} f_X(x)dx = 1$, then $f_X(x)$ is called the *probability density function (p.d.f.)* of X
- A random variable X is continuous if there exists a p.d.f. of X
- Support of X: $\{x \in \mathbb{R} \mid f_X(x) > 0\}$
- C.d.f. and p.d.f.:
$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy$$
$$f_X(x) = F'_X(x) \text{ for any } x \text{ at which } F_X \text{ is differentiable}$$
- Types of r.v.:
  1. Discrete: Support is countable
  2. Continuous: P.d.f. exists $\Rightarrow$ support is uncountable
  3. Neither: Support is uncountable but p.d.f. does not exist

Uniform Distribution

- "Completely random" number over a continuous interval
    - Nigeria survey: Direction a stick falls
- X follows the Uniform distribution on the interval $[a, b]$:

$$F_X(x; a, b) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & b < x \end{cases}$$

$$f_X(x; a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Parameters of the Uniform distribution: a and b
- Denoted by: $X \sim \text{Unif}(a, b)$
- Commonly on the unit interval $[0, 1]$

- Not many real-world examples, unless artificially created
- Useful tool for modeling and simulations

Functions of Random Variables

- How to generate random numbers?
  - Online survey: Randomly switch questions
  - Experiment: Randomly assign treatment or control
- Function of r.v. is also r.v.:
  - If $X : \Omega \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, then $g \circ X : \Omega \to \mathbb{R}$
  - $Y(\omega) \equiv g(X(\omega))$ is r.v.
- If X is discrete:
  - P.f. of Y: $f_Y(y) \equiv \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \sum_{\{x \mid g(x) = y\}} f_X(x)$
  - E.g., $Y = n - X$ where $X \sim \text{Binom}(n, p)$: $f_Y(y) = \binom{n}{n-y} p^{n-y}(1-p)^y$
- If X is continuous:
  - C.d.f. of Y:

$$F_Y(y) \equiv \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \in \{x \mid g(x) \leq y\})$$
$$= \int_{\{x \mid g(x) \leq y\}} f_X(x) dx$$

  - E.g., $Y = X^2$ where $X \sim \text{Unif}(0, 1)$: $F_Y(y) = \int_0^{\sqrt{y}} 1 dx$

Inverse-CDF Method

- Increasing function of r.v.:
  - $g : \mathbb{R} \to \mathbb{R}$ is increasing: $a < b$ implies that $g(a) < g(b)$
  - If g is increasing, then $F_Y(y) = F_X(g^{-1}(y))$

- Quantile function: The *quantile function (q.f.)* (a.k.a. inverse c.d.f.) of X, denoted by $Q_X : [0, 1] \to \mathbb{R}$, is a function
$$Q_X(u) \equiv \inf\{x \mid F_X(x) > u\}$$

- Inverse-CDF method: Let X is an r.v. with $Q_X(\cdot)$, $U \sim \text{Unif}(0, 1)$, and $Y = Q_X(U)$. Then, $F_Y(y) = F_X(y)$
  - $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Q_X(U) \leq y) = \mathbb{P}(U \leq F_X(y)) = F_X(y)$
  - $\inf\{x \mid F_X(x) > U\} \leq y \Leftrightarrow U \leq F_X(y)$: **Proof.** Samuel's section
- Generating random numbers whose c.d.f. is F_X:
  - Generate $U \sim \text{Unif}(0, 1)$
  - Transform by $X = Q_X(U)$
- Special Case:
  - F_X is increasing $\Rightarrow Q_X(U) = F_X^{-1}(U)$
  - $F_U(Q_X^{-1}(x)) = F_U((F_X^{-1})^{-1}(x)) = F_U(F_X(x)) = F_X(x)$

Multivariate Random Variables

- Multiple r.v.s:
  - Dice roll: Indicator of each number on the dice
  - Survey: Responses by many respondents

- Joint c.d.f.: The *joint c.d.f.* of a random vector $\mathrm{X} \equiv (X_1, \ldots, X_n)$, denoted by $F_{\mathrm{X}} : \mathbb{R}^n \to [0, 1]$, is a function
$$F_{\mathrm{X}}(\mathrm{x}) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$
where $\mathrm{x} \equiv (x_1, \ldots, x_n)$
- Dice roll and indicator: $X_i \equiv 1\{i \text{ shows on the dice}\}$
  - X_i is either 0 or 1
  - $F_{\mathrm{X}}(\mathrm{x}) = j/6$ where j is the number of 1 in x

- Joint p.f.: Let $X_1, \ldots, X_n$ be discrete r.v.s. The *joint p.f.* of X, denoted by $f_{\mathrm{X}} : \mathbb{R}^n \to [0, 1]$ is a function
$$f_{\mathrm{X}}(\mathrm{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$
- Dice roll and indicator, again:
  - $f_{\mathrm{X}}(\mathrm{x}) = 1/6$ for any x such that only one element is 1
  - $f_{\mathrm{X}}(\mathrm{x}) = 0$ otherwise

Multinomial Distribution

- The number of times each "category" appears in n trials
  - Multiple dice rolls: How many times each number shows
  - Responses to multiple choice/count questions
  - Word counts in a document (Problem Set 3)

- X follows a Multinomial distribution:
  - Joint p.f.: For non-negative integers $x_1, \dots x_K$,

$$f_{\mathrm{X}}(\mathrm{x}; n, \mathrm{p}) = \begin{cases} \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise} \end{cases}$$

  - Parameters of the Multinomial distribution: n and p
  - $p_1 + \cdots + p_k = 1$
  - Denoted by: $\mathrm{X} \sim \mathrm{Multi}(n, \mathrm{p})$

- Multinomial is Binomial if $k = 2$
- Multinomial is the sum of indicators for each trial

Multivariate Uniform

- Joint p.d.f.: For a random vector $\mathrm{X} = (X_1, \ldots, X_n)$, if there exists a function $f_{\mathrm{X}} : \mathbb{R}^n \to \mathbb{R}^+$ such that for any set $C \subset \mathbb{R}^n$,
$$\mathbb{P}(\mathrm{X} \in C) = \int \underset{C}{\cdots} \int f_{\mathrm{X}}(\mathrm{x}) dx_1 \ldots dx_n$$
and $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathrm{X}}(\mathrm{x}) dx_1 \ldots dx_n = 1$, then $f_{\mathrm{X}}(\mathrm{x})$ is called the *joint p.d.f.* of X

- $\mathrm{X} = (X_1, X_2)$ follows a Uniform distribution over $[0, 1] \times [0, 1]$:
    - Joint p.d.f.:
    $$f_{\mathrm{X}}(\mathrm{x}) = \begin{cases} 1 & 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
    - Joint c.d.f.:
    $$F_{\mathrm{X}}(\mathrm{x}) = \begin{cases} 0 & x_1 < 0 \text{ or } x_2 < 0 \\ x_1 x_2 & 0 \leq x_1 \leq 1 \text{ and } 0 \leq x_2 \leq 1 \\ 1 & x_1 > 1 \text{ and } x_2 > 1 \end{cases}$$

Marginal Distribution

- Marginal c.d.f.: Let $\mathrm{X} = (X_1, \ldots, X_n)$ be a random vector and F_{X} be its joint c.d.f. Then,
  $F_{X_i}(x_i) = \lim_{x_1 \to \infty} \ldots \lim_{x_{i-1} \to \infty} \lim_{x_{i+1} \to \infty} \ldots \lim_{x_n \to \infty} F_{\mathrm{X}}(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots x_n)$
  F_{X_i} is called the *marginal c.d.f.* of X_i: **Proof**. $n = 2$ case
- F_{X_i} is a valid c.d.f.: The *marginal distribution* of X_i

- Marginal p.f.: If the marginal distribution of X_i is discrete, the *marginal p.f.* of X_i is defined by
  $$f_{X_i}(x) \equiv \mathbb{P}(X_i = x_i) = F_{X_i}(x_i) - \lim_{y \uparrow x_i} F_{X_i}(y)$$
- If a joint p.f. $f_{\mathrm{X}}(\mathrm{x})$ exists, $f_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} f_{\mathrm{X}}(\mathrm{x})$

- Marginal p.d.f.: If the marginal c.d.f. F_{X_i} has a p.d.f. f_{X_i}, it is called a *marginal p.d.f.* of X_i
- If a joint p.d.f. $f_{\mathrm{X}}(\mathrm{x})$ exists,
  $f_{X_i}(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_n} f_{\mathrm{X}}(\mathrm{x}) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$

Independence

- Independent r.v.s: **Very** important in data analysis
- Independence of r.v.s: R.v.s $X_1, \dots, X_n$ are *independent* if and only if for any subsets $C_1, \dots C_n$ of $\mathbb{R}$,
$$\mathbb{P}(X_1 \in C_1, \dots, X_n \in C_n) = \mathbb{P}(X_1 \in C_1) \dots \mathbb{P}(X_n \in C_n)$$
- Notation: $X_i \perp\!\!\!\perp X_j$
- Knowing the value of X_i does not help predict X_j

- Connection with marginal distribution:
  - $X_1, \dots, X_n$ are independent if and only if for any $x_1, \dots, x_n \in \mathbb{R}$
$$F_{\mathrm{X}}(\mathrm{x}) = \prod_{i=1}^{n} F_{X_i}(x_i)$$
  - If a joint p.f. $f_{\mathrm{X}}(\mathrm{x})$ exists, $f_{\mathrm{X}}(\mathrm{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$
  - If a joint p.d.f. $f_{\mathrm{X}}(\mathrm{x})$ exists, $f_{\mathrm{X}}(x) = \prod_{i=1}^{n} f_{X_i}(x_i)$

I.i.d. and Random Sample

- Benchmark model of data generating process
  1. Data points are independent random variables
  2. All data points follow a common distribution
- Independent and identically distributed: $X_1, \ldots, X_n$ are *i.i.d. (independent and identically distributed)* if and only if they are independent and each has the same marginal distribution with c.d.f. F
- Notation: $X_i \overset{\text{i.i.d.}}{\sim} F$ or $X_i \overset{\text{i.i.d.}}{\sim} f$
- $(X_1, \ldots, X_n)$ is called *a random sample of size n from F*

- Random sampling for opinion poll:
  - Population N, Dem supporters $m_D < N$, $p \equiv m_D/N$
  - X_i: 1 if person i is Dem supporter, 0 otherwise
  - i is randomly chosen:
    1. X_i is i.i.d. Bernoulli with p with the *super population* assumption
    2. X_i is independent but not indentically distributed under the *finite population* assumption

Convolution

- Sum of two random variables is called *convolution*
- Convolution: Let X_1 and X_2 be independent random variables and $Y \equiv X_1 + X_2$. Then, the distribution of Y is called the *convolution* of the distributions of X_1 and X_2.
- If both X_1 and X_2 are discrete,
$$f_Y(y) = \sum_{x_1=-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1)$$
- If both X_1 and X_2 are continuous,
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1) dx_1$$
- Convolution of Bernoulli r.v.s with common p is $\text{Binom}(2, p)$:
$$f_Y(y) = \sum_{x_1=0}^{1} f_{X_1}(x_1) f_{X_2}(y - x_1) = \begin{cases} (1-p)^2 & (y = 0) \\ 2p(1-p) & (y = 1) \\ p^2 & (y = 2) \end{cases}$$

Conditional Distribution

- In many studies, *prediction* is of interest
  - Vote choice given ethnicity, gender, age, etc...
  - Attitude toward immigration given occupation
  - Economic growth/conflict behavior given regime type
- *Regression* (covered in 699) is the most important method
- *Conditional distribution* is the idea behind regression
- Conditional c.d.f.: Let $\mathrm{X} \equiv (X_1, \ldots, X_n)$ have the joint distribution with c.d.f. F_{X}. Then,

$$\begin{aligned}
&F_{\mathrm{X}_{1:i}|\mathrm{X}_{(i+1):n}}(x_1, \ldots, x_i \mid \mathrm{X}_{(i+1):n} \in \times_{j=i+1}^{n} C_j) \\
\equiv&\mathbb{P}(X_1 \leq x_1, \ldots, X_i \leq x_i \mid X_{i+1} \in C_{i+1}, \ldots, X_n \in C_n) \\
=&\frac{F_{\mathrm{X}}(x_1, \ldots, x_n)}{\mathbb{P}(X_{i+1} \in C_{i+1}, \ldots, X_n \in C_n)}
\end{aligned}$$

  for $C_j \subset \mathbb{R}, j = i+1, \ldots, n$, is called the *conditional c.d.f. of* $\mathrm{X}_{1:i}$ *given that* $X_{i+1} \in C_{i+1}, \ldots, X_n \in C_n$
- Conditional c.d.f. uniquely defines the *conditional distribution of* $\mathrm{X}_{1:i}$ *given that* $\mathrm{X}_{(i+1):n} \in \times_{j=i+1}^{n} C_j$

Conditional P.(d.)f. and Hybrid Random Vectors

- If X is discrete, then there exists the *conditional p.f.*:
$$f_{X_{1:i}|X_{(i+1):n}}(x_1, \ldots, x_i \mid x_{i+1}, \ldots, x_n) = \frac{f_X(x_1, \ldots, x_n)}{f_{X_{(i+1):n}}(x_{i+1}, \ldots, x_n)}$$
- If X is continuous, then there exists the *conditional p.d.f.*:
$$f_{X_{1:i}|X_{(i+1):n}}(x_1, \ldots, x_i \mid x_{i+1}, \ldots, x_n) = \frac{f_X(x_1, \ldots, x_n)}{f_{X_{(i+1):n}}(x_{i+1}, \ldots, x_n)}$$
- Independence $\Leftrightarrow f_{X_1|X_2}(x_1 \mid x_2) = f_{X_1}(x_1)$
- Joint p.(d.)f. = cond. p.(d.)f. $\times$ marg. p.(d.)f.
- Joint p.f.-p.d.f. of a hybrid random vector: Let $X_{(i+1):n}$ have a marginal p.d.f. (p.f.) $f_{X_{(i+1):n}}$ and $X_{1:i}$ have the conditional p.f. (p.d.f.) $f_{X_{1:i}|X_{(i+1):n}}$. Then, we define the *joint p.f.-p.d.f.* of X as
$$\begin{aligned} &f_X(x_1, \ldots, x_n) \\ \equiv &f_{X_{1:i}|X_{(i+1):n}}(x_1, \ldots, x_i \mid x_{i+1}, \ldots, x_n) f_{X_{(i+1):n}}(x_{i+1}, \ldots, x_n) \end{aligned}$$

Uniform-Binomial Model

- A popular model in Bayesian statistics:
  1. Proportion of Dem supporters drawn from the Uniform
  2. Random sample of size n for a survey on partisanship
- Data generating process:
  1. Proportion of Dem supporters: $X_1 \sim U[0,1]$
  2. Number of Dem supporters in sample: $X_2 \sim \text{Binom}(n, X_1)$
- Joint p.f.-p.d.f.:

$$f_{X_1,X_2}(x_1,x_2) = f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1) = 1 \times \binom{n}{x_2} x_1^{x_2}(1-x_1)^{n-x_2}$$

- Marginal p.d.f. and p.f.:
  1. Marginal p.d.f. of X_1:

$$f_{X_1}(x_1) = \begin{cases} 1 & (0 \leq x_1 \leq 1) \\ 0 & (\text{otherwise}) \end{cases}$$

  2. Marginal p.f. of X_2: Letting $B(\cdot,\cdot)$ be the Beta function

$$f_{X_2}(x_2) = \binom{n}{x_2} \int_0^1 x_1^{x_2}(1-x_1)^{n-x_2} dx_1 = \binom{n}{x_2} B(x_2+1, n-x_2+1)$$

Bayes' Theorem for Random Variables

- The number of Dem supporters in sample is known
- Need to estimate the proportion of Dems in population

- Bayes' theorem for r.v.s: Let (X_1, X_2) has a joint p.f., p.d.f., or p.f.-p.d.f. Then,

$$f_{X_1}(x_1) = \frac{f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1)}{f_{X_2}(x_2)}$$

- Law of total probability for r.v.s: Let (X_1, X_2) has a joint p.f., p.d.f., or p.f.-p.d.f. Then,

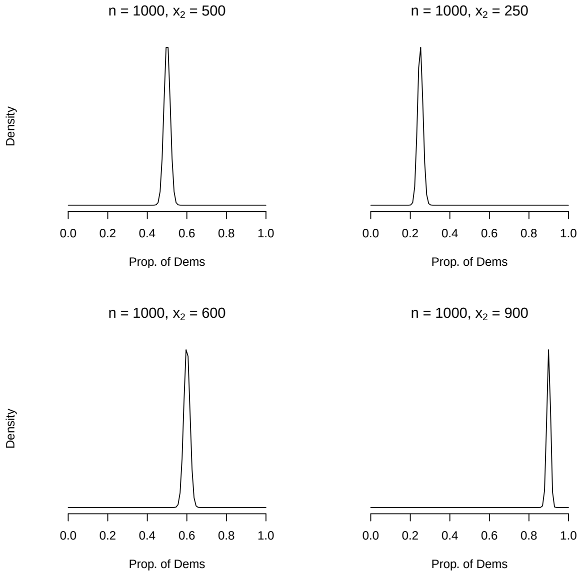
$$f_{X_2}(x_2) = \sum_{x_1} f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1) \quad (X_1 \text{ is discrete})$$

$$f_{X_2}(x_2) = \int_{x_1} f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1) dx_1 \quad (X_1 \text{ is continuous})$$

- *Posterior* p.d.f. of X_1 given X_2 in the Uniform-Binomial:

$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{x_1^{x_2}(1 - x_2)^{n - x_2}}{B(x_2 + 1, n - x_2 + 1)}$$

Posterior P.d.f.



Expectation

- **BH**, Ch. 4 and p. 200; **DS**, Ch. 4; **W**, Ch. 3; **CB**, 2.2
- Summary of r.v. X:
  - What is the gain you expect from a lottery?
  - What is the number of Dems you expect in a sample?
  - What is the lifetime income you expect from an academic job?

- Expectation or expected value of X: Weighted average of X where the weights are the probability measure of events $X = x$
- For a discrete r.v. X,

$$\mathbb{E}[X] = \sum_x x f_X(x)$$

- For a continuous r.v. X,

$$\mathbb{E}[X] = \int_x x f_X(x) dx$$

- $X \sim \text{Bern}(p) \Rightarrow \mathbb{E}[X] = p$
- $X \sim \text{Unif}(0, 1) \Rightarrow \mathbb{E}[X] = 1/2$

Existence of Expectation

- Expectation does not always exist
- Existence of expectation: $\mathbb{E}[X]$ exists if and only if $\mathbb{E}[X_-] < \infty$ or $\mathbb{E}[X_+] < \infty$, where $X_- \equiv -\min\{X, 0\}$ and $X_+ \equiv \max\{X, 0\}$.
  1. $\mathbb{E}[X_-] < \infty$ and $\mathbb{E}[X_+] < \infty$: $-\infty < \mathbb{E}[X] < \infty$
  2. $\mathbb{E}[X_-] < \infty$ and $\mathbb{E}[X_-] = \infty$: $\mathbb{E}[X] = -\infty$
  3. $\mathbb{E}[X_+] = \infty$ and $\mathbb{E}[X_-] < \infty$: $\mathbb{E}[X] = \infty$
  4. $\mathbb{E}[X_+] = \infty$ and $\mathbb{E}[X_-] = \infty$: $\mathbb{E}[X]$ does not exist
- Expectation can be infinity, but its sign should be well defined

- X follows the standard Cauchy distribution:

$$f_X(x) = \frac{1}{\pi(1 + x^2)} \text{ for } -\infty < x < \infty$$

  - Valid p.d.f: $\int_{-\infty}^{\infty} f(x)dx = [\tan^{-1}(x)/\pi]_{-\infty}^{\infty} = \{\pi/2 - (-\pi/2)\}/\pi = 1$
  - $\int_0^{\infty} xf(x)dx = [\log(1 + x^2)/2]_0^{\infty} = \infty$
  - Similarly, $\int_{-\infty}^{0} -xf(x)dx = [\log(1 + x^2)/2]_{-\infty}^{0} = \infty$
  - Expectation does not exist for the Cauchy distribution

Indicator and Linearity

- Expectation of Indicator: For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let A be an event and define r.v. $1_A \equiv 1\{\omega \in A\}$. Then, $\mathbb{E}[1_A] = \mathbb{P}(A)$
- Corollary: Let $C \subset \mathbb{R}$. For r.v. X, define $1_C(X) \equiv 1\{X \in C\}$. Then, $\mathbb{E}[1_C(X)] = \mathbb{P}(X \in C)$
- Dice roll: X is the number on the face
    - Let $C \equiv \{2, 3, 4, 5\}$
    - $\mathbb{P}(X \in C) = 2/3$
    - $\mathbb{E}[1_C(X)] = 1 \times (4 \times 1/6) + 0 \times (2 \times 1/6) = 2/3$

- Linearity: Let X_1, X_2 be r.v.s. Then,
$$\mathbb{E}[aX_1 + bX_2 + c] = a\mathbb{E}[X_1] + b\mathbb{E}[X_2] + c$$
- Binomial expectation:
    - By definition of expactation,
$$\mathbb{E}[X] = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} = np$$
    - $\text{Binom}(n, p)$ is the distribution of $\sum_{i=1}^{n} X_i$ where $X_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$
$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np$$

Random Vectors

- Expectation of random vector: For a random vector $\mathbf{X}$, its expectation is defined as
$$\mathbb{E}[\mathbf{X}] \equiv (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])$$
where the expectation of X_i is over its marginal distribution

- Multinomial distribution $\mathbf{X} \sim \text{Multi}(n, \mathbf{p})$:
  1. Marginal distribution of X_1 is Binomial:
$$\begin{aligned}\mathbb{P}(X_1 = x_1) &= \sum_{x_2 \ldots x_K} \frac{n!}{x_1! \ldots x_K!} p_1^{x_1} \ldots p_K^{x_K} \\ &= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} \sum_{x_2 \ldots x_K} \frac{(n-x_1)!}{x_2! \ldots x_K!} p_2^{x_2} \ldots p_K^{x_K} \\ &= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} (1-p_1)^{n-x_1}\end{aligned}$$
  2. $\mathbb{E}[\mathbf{X}] = (np_1, \ldots, np_K)$

Functions and Product

- Expectation of functions of r.v.: Let X be a r.v. and $g : \mathbb{R} \to \mathbb{R}$. Then,

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) f_X(x) & (X \text{ discrete}) \\ \int_x g(x) f_X(x) dx & (X \text{ continuous}) \end{cases}$$

  Proof. Directly follows from the fact that for any $C \subset \mathbb{R}$, $\mathbb{P}(g(X) \in C) = \mathbb{P}(X \in \{x \in \mathbb{R} \mid g(x) \in C\})$

- X follows a Geometric distribution, $X \sim \text{Geom}(p)$:

$$f_X(x) = (1-p)^{x-1} p, \text{ for } x = 1, 2, \ldots$$

  1. St. Petersburg paradox: $g(x) \equiv 2^x \Rightarrow \mathbb{E}[g(X)] = \infty$ if $p = 1/2$
  2. $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ in general: $\mathbb{E}[X] = 2$

- Lemma: Let X be a discrete r.v. whose support is the non-negative integers. Then, $\mathbb{E}[X] = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)$

- Product of independent r.v.s: Let $X_i, i = 1, \ldots, n$ are independent. Then, $\mathbb{E}[\prod_{i=1}^{n} X_n] = \prod_{i=1}^{n} \mathbb{E}[X_i]$
- $X_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p) \Rightarrow \mathbb{P}(X_1 = 1, \ldots, X_n = 1) = p^n$

Inequalities of Expectation

- If $X_1 \leq X_2$ with probability 1, i.e., $X_1(\omega) \leq X_2(\omega)$ for all $\omega \in \Omega$, then $\mathbb{E}[X_1] \leq \mathbb{E}[X_2]$
- If $a \leq X \leq b$ with probability 1, i.e., $a \leq X(\omega) \leq b$ for all $\omega \in \Omega$, then $a \leq \mathbb{E}[X] \leq b$

- Jensen's inequality: Let $g : \mathbb{R} \to \mathbb{R}$ be a concave (convex) function. Then, for a random vector X, $\mathbb{E}[g(\mathrm{X})] \leq (\geq) g(\mathbb{E}[\mathrm{X}])$
- Concave function: A function $g : \mathbb{R}^n \to \mathbb{R}$ is concave if and only if for every $a \in (0, 1)$,
$$g\big(a\mathrm{x} + (1-a)\mathrm{y}\big) \geq ag(\mathrm{x}) + (1-a)g(\mathrm{y})$$
for any $x, y \in \mathbb{R}^n$
- Logarithm is common in statistics: $\mathbb{E}[\log(X)] < \log\left(\mathbb{E}[X]\right)$

Moments and Variance

- Moments of an r.v.: For an r.v. X and a positive integer k, $\mathbb{E}[X^k]$ is called the kth *moment* of X
- Existence of moments: If $\mathbb{E}[X^k]$ exists, $\mathbb{E}[X^l]$ exists for any $l < k$
- Central moments: $\mathbb{E}[(X - \mathbb{E}[X])^k]$ is called the kth *central moment* or the kth *moment of X about the mean*
- If the kth moment exists, the lth central moment exists for $l \leq k$

- Variance: The second centeral moment of X is called the *variance* of X, denoted by $\mathbb{V}(X) \equiv \mathbb{E}[(X - \mathbb{E}[X])^2]$
- Variance and moments: $\mathbb{V}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- $\mathbb{V}(X) \geq 0$, with equality if and only if $\mathbb{P}(X = c) = 1$ for some c
- $Y = aX + b \Rightarrow \mathbb{V}(Y) = a^2\mathbb{V}(X)$
- Variance of Bern(p): $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p)$

Covariance

- Covariance: For r.v.s X_1 and X_2, the *covariance* of X_1 and X_2, denoted by $\text{Cov}(X_1, X_2)$, is defined as:
$$\text{Cov}(X_1, X_2) \equiv \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]$$
- Analogously to the variance, $\text{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$
- Correlation: The *correlation* of X_1 and X_2, denoted by $\rho(X_1, X_2)$, is defined as:
$$\rho(X_1, X_2) \equiv \frac{\text{Cov}(X_1, X_2)}{\sqrt{\mathbb{V}(X_1)\mathbb{V}(X_2)}}$$
- $\rho(X_1, X_2) = \text{Cov}\left((X_1 - \mathbb{E}[X_1])/\sqrt{\mathbb{V}(X_1)}, (X_2 - \mathbb{E}[X_2])/\sqrt{\mathbb{V}(X_2)}\right)$
- Covariance depends on the scale of r.v.s, but $|\rho(X_1, X_2)| \leq 1$

- X_1 and X_2 are *uncorrelated* if and only if $\text{Cov}(X_1, X_2) = 0$
- X_1 and X_2 are independent $\Rightarrow$ X_1 and X_2 are uncorrelated
- The converse does not necessarily hold:
    - $U \sim \text{Unif}(0, 1)$, $X_1 = \cos 2\pi U$ and $X_2 = \sin 2\pi U$
    - Clearly, X_1 and X_2 are not independent, but $\text{Cov}(X_1, X_2) = 0$
- Covariance and correlation indicate *linear* relationship b/w r.v.s

Variance-Covariance Matrix

- Trivially, $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])] = \text{Cov}(X, X)$
- Variance-covariance Matrix: For r.v.s $X_1, \ldots, X_n$, we define the *(variance-)covariance matrix*, denoted by $\mathbb{V}(\mathrm{X})$ or Σ_{X}, as

$$\Sigma_{\mathrm{X}} \equiv \begin{pmatrix} \mathbb{V}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \mathbb{V}(X_n) \end{pmatrix}$$

- In vector notation, $\Sigma_{\mathrm{X}} = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])(\mathrm{X} - \mathbb{E}[\mathrm{X}])^\top]$
- Σ_{X} is positive semi-definite
- Σ_{X} is positive definite unless some r.v.s are constant $\Rightarrow$ invertible

- If $X_1, \ldots, X_n$ are i.i.d., Σ_{X} is diagonal
- $\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i) + 2\sum_{i<j} a_i a_j \text{Cov}(X_i, X_j)$
- X_is are uncorrelated, $\mathbb{V}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{V}(X_i)$
- If $X_1, \ldots, X_n$ are i.i.d., $\mathbb{V}\left(\sum_{i=1}^n X_i\right) = \text{tr}(\Sigma_{\mathrm{X}})$

Conditional Expectation

- Summarizing prediction of Y using X:
  - Given $X = x$, Y is predicted by its conditional *distribution*
  - Conditional *expectation* is used to summarize the prediction
- Conditional expectation: The *conditional expectation* of Y given X, denoted by $\mathbb{E}[Y \mid X]$, is the expectation of the conditional distribution of Y given X
- Conditional moments are defined with expectation replaced by conditional expectation, e.g. $\mathbb{V}(Y \mid X) \equiv \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2 \mid X]$
- *Conditional expectation is an r.v.*:
  - $\mathbb{E}[Y \mid X]$ is a function of X
  - $\mathbb{E}[Y \mid X]$ has a distribution defined by F_X
- *Conditional expectation is expectation*:
  - For any fixed x_1, $\mathbb{E}[Y \mid X = x_1]$ is expectation
  - All the properties of expectation hold for $\mathbb{E}[Y \mid X = x_1]$
- Uniform-Binomial example:

$$\mathbb{E}[X_1 \mid X_2] = \int_0^1 x_1 \frac{x_1^{X_2}(1 - x_1)^{n - X_2}}{B(X_2 + 1, n - X_2 + 1)} dx_1 = \frac{X_2 + 1}{n + 2}$$

Properties of Conditional Expectation

- Law of iterated expectations: Let X and Y be r.v.s. Then,
$$\mathbb{E}\left[\mathbb{E}[g(X,Y) \mid X]\right] = \mathbb{E}[g(X,Y)]$$
for any function g.
- Law of total variance: Let X and Y be r.v.s. Then,
$$\mathbb{E}\left[\mathbb{V}(Y \mid X)\right] + \mathbb{V}\left(\mathbb{E}[Y \mid X]\right) = \mathbb{V}(Y)$$
- Uniform-Binomial example:
$$\mathbb{E}[X_2] = \mathbb{E}[nX_1] = \frac{n}{2}$$
$$\mathbb{V}[X_2] = \mathbb{E}[nX_1(1 - X_1)] + \mathbb{V}(nX_1) = \frac{n}{6} + \frac{n^2}{12}$$
- Minimization of expected squared distance (regression): Conditional expectation is the "best" predictor in the sense that
$$\operatorname*{argmin}_c \mathbb{E}\left[(Y - c)^2 \mid X\right] = \mathbb{E}[Y \mid X]$$

Proof.
$$\mathbb{E}\left[(Y - c)^2 \mid X\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2 \mid X\right] + \underbrace{(\mathbb{E}[Y \mid X] - c)^2}_{=0 \text{ iff } c = \mathbb{E}[Y|X]}$$

Standard Gaussian Distribution

- X follows the standard multivariate Gaussian distribution:
  - Joint p.d.f.: For a vector of real numbers $\mathrm{x} \equiv x_1 \dots, x_K$,

$$f_{\mathrm{X}}(\mathrm{x}) = \frac{1}{(2\pi)^{K/2}} e^{-\frac{1}{2}\mathrm{x}^\top \mathrm{x}} = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_k^2}{2}}$$

  - Denoted by: $\mathrm{X} \sim \mathcal{N}(0, \mathbf{I}_K)$
  - $X_1, \dots, X_K$ are independent
  - $K = 1$: The standard Gaussian (Normal) distribution $\mathcal{N}(0, 1)$

- Box-Muller Transfromation: Let U_1 and U_2 be independent uniform r.v.s. Define

$$X_1 = \sqrt{-2\log U_1}\cos(2\pi U_2),$$
$$X_2 = \sqrt{-2\log U_1}\sin(2\pi U_2).$$

  Then,

$$\mathrm{X} \sim \mathcal{N}(0, \mathbf{I}_2)$$

- Random number generator for the Gaussian distributions

Change of Variables

- **BH**, 8.1; **DS**, p. 172–3, 182–6
- Change of variables: Let X be a continuous random vector of length K and Y $\equiv g(X)$ where $g : \mathbb{R}^K \to \mathbb{R}^K$ is one-to-one and differentiable. Then,

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det \left(J_{g^{-1}}(y) \right) \right|$$

Proof (univariate version). Appendix slides.

- The inverse of the Box-Muller transformation is:

$$U_1 = e^{-(X_1^2+X_2^2)/2}$$

$$U_2 = \frac{1}{2\pi} \arctan\left(\frac{X_2}{X_1}\right)$$

- The determinant of the Jacobian is:

$$\det\left(J_{g^{-1}}(y)\right) = \det\begin{pmatrix} \frac{\partial U_1}{\partial X_1} & \frac{\partial U_1}{\partial X_2} \\ \frac{\partial U_2}{\partial X_1} & \frac{\partial U_2}{\partial X_2} \end{pmatrix} = -\frac{1}{\sqrt{2\pi}} e^{-X_1^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-X_2^2/2}$$

Linear Transformation of Gaussian

- *Linear transformation*: Let X be a K-dimensional random vector. A linear transformation of X is
$$\mathrm{Y} = \mathrm{a} + \mathbf{A}\mathrm{X}$$
where $\mathbf{A}$ is a matrix with K columns
- Multivariate Gaussian: Let X follow the standard multivariate Gaussian distribution. Then, for a full rank $K \times K$ matrix $\mathbf{A}$ and an K dimensional vector μ, $\mathrm{Y} = \mu + \mathbf{A}\mathrm{X}$ has a p.d.f.:
$$f_{\mathrm{Y}}(\mathrm{y}) = \frac{1}{(2\pi)^{K/2}\det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathrm{x}-\mu)^{\top}\Sigma^{-1}(\mathrm{x}-\mu)}$$
where $\Sigma = \mathbf{A}\mathbf{A}^{\top}$
Proof.
$$\mathrm{X} = \mathbf{A}^{-1}(\mathrm{Y} - \mu)$$
$$J = \mathbf{A}^{-1}$$
- $\mathbb{E}[\mathrm{Y}] = \mu$ and $\mathbb{V}(\mathrm{Y}) = \Sigma$
- Uncorrelated $\Leftrightarrow$ (pairwise) independent

Conditional and Marginal of Gaussian

- Let $(X_1, X_2)^\top \sim \mathcal{N}(\mu, \Sigma)$. The joint p.d.f. is:

$$f_{(X_1,X_2)}(x_1,x_2) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$= \underbrace{\frac{e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(x_2-\mu_2-\rho\sigma_2\frac{x_1-\mu_1}{\sigma_1}\right)^2}}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}}_{f_{X_2|X_1}(x_2|x_1)} \underbrace{\frac{e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}}{\sqrt{2\pi}\sigma_1}}_{f_{X_1}(x_1)}$$

- $X_1 \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), X_2 \mid X_1 \sim \mathcal{N}\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(X_1 - \mu_1), (1-\rho^2)\sigma_2^2\right)$
- Both marginal and conditional distributions are Gaussian

- Regression $\mathbb{E}[X_2 \mid X_1] = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(X_1 - \mu_1)$ is linear in X_1
  $\rightsquigarrow$ linear regression

Moment Generating Function

- Moment generating function: Let X be an r.v. The *moment generating function (m.g.f.)* of X, denoted by $M_X(t)$, is defined as
$$M_X(t) = \mathbb{E}[e^{tX}]$$
if $\mathbb{E}[e^{tX}]$ exists for all $t \in (-s, s)$ for some $s > 0$.

- If m.g.f. is given, higher order moments can be easily computed:
$$\left.\frac{d^k}{dt^k}M_X(t)\right|_{t=0} = \mathbb{E}[X^k e^{0X}] = \mathbb{E}[X^k], \text{ for } k = 1, 2, \ldots$$

- If $X_1, \ldots, X_n$ are independent, the m.g.f. of $Y \equiv \sum_{i=1}^n X_i$ is
$$M_Y(t) = \mathbb{E}[e^{t\sum_{i=1}^n X_i}] = \mathbb{E}[\prod_{i=1}^n e^{tX_i}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \prod_{i=1}^n M_{X_i}(t)$$

- M.g.f. uniquely determines the distribution: If r.v.s X_1 and X_2 have m.g.f.s and $M_{X_1}(t) = M_{X_2}(t)$ for all $t \in (-a, a)$ for some $a > 0$, then c.d.f.s $F_{X_1}(x) = F_{X_2}(x)$ for all x

M.g.f. of Gamma Distributions

- Square of a Gaussian r.v. follows a *Gamma distribution* (PS9, Q4)
- X follows a Gamma distribution:
  - P.d.f.:
    $$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for} \quad x > 0$$
  - Parameters: Shape $\alpha > 0$ and rate $\beta > 0$
  - Alternative parameterization: Shape $\alpha > 0$ and scale $\theta = 1/\beta$
  - Gamma function: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$
  - Denoted by: $X \sim \text{Ga}(\alpha, \beta)$
- M.g.f. of the Gamma distribution:
  $$\mathbb{E}[e^{tX}] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} = \left(\frac{\beta}{\beta-t}\right)^\alpha$$
- Expectation and variance: $\mathbb{E}[X] = \alpha/\beta, \mathbb{V}[X] = \alpha/\beta^2$
- Sum of independent Gamma r.v.s $X_i \sim \text{Ga}(\alpha_i, \beta), i = 1, \ldots, n$:
  $$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \left(\frac{\beta}{\beta-t}\right)^{\alpha_i} = \left(\frac{\beta}{\beta-t}\right)^{\sum_{i=1}^n \alpha_i}$$
  $\Rightarrow \sum_{i=1}^n X_i \sim \text{Ga}(\sum_{i=1}^n \alpha_i, \beta)$

Sample Moments

- Let $X_1, \ldots, X_n$ is a random sample from a distribution F_X
- In other words, $X_1, \ldots, X_n$ are *data*
- Sample moments: Let $X_1, \ldots, X_n$ be i.i.d. r.v.s. The kth *sample moment*, denoted by M_k, is defined as
$$M_k \equiv \frac{1}{n} \sum_{i=1}^{n} X_i^k$$
- Why this is important: We never observe F_X, hence neither $\mathbb{E}[X^k]$
- We use M_k as an *estimator*–a function of r.v.s, therefore r.v.
- Mean and variance of M_k for any F_X:
$$\mathbb{E}[M_k] = \mathbb{E}[X^k], \quad \mathbb{V}(M_k) = \frac{\mathbb{V}(X^k)}{n}$$
- In particular, for sample mean $\bar{X} \equiv \sum_{i=1}^{n} X_i / n$,
$$\mathbb{E}[\bar{X}] = \mathbb{E}[X], \quad \mathbb{V}(\bar{X}) = \frac{\mathbb{V}(X)}{n}$$
- If we specify F_X, we can derive the distribution of M_k

Sample Mean of Gaussian R.v.

- We want to find the distribution of the sum of independent r.v.s $\rightsquigarrow$ use m.g.f.!
- M.g.f. of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$:

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

- M.g.f. of the sample mean:

$$\mathbb{E}[e^{t\overline{X}}] = \prod_{i=1}^{n} \mathbb{E}[e^{\frac{t}{n}X}] = e^{n\left\{\mu\frac{t}{n}+\frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right\}} = e^{\mu t+\frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2}$$

$$\Rightarrow \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$