50.034 - Introduction to Probability and Statistics

Week 8 – Lecture 14 (Review Lecture)

January–May Term, 2019

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Lecture 14 Assigned Readings: Nil

Outline of Lecture

Review of Ten Topics:

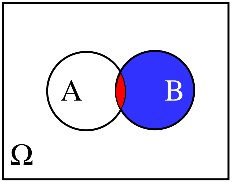
- Conditional probability and independent events
- Law of total probability, Bayes' theorem
- Probability/joint/conditional distributions
- Joint/marginal pmf/pdf/cdf
- Expectation and variance
- Covariance, correlation, independence
- Moments, moment generating functions
- Special distributions (Poisson, exponential, etc.)
- Markov's inequality, Chebyshev's inequality
- Law of large numbers, central limit theorem

Conditional probability

Recall: An event is a subset of outcomes in a sample space Ω.

For any two events A and B with $\Pr(B) > 0$, the conditional probability of A given B is defined by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$



Interpretation:
Since event B has occurred, the relevant outcomes are no longer all possible outcomes in the sample space, but only those outcomes that are contained in B. So event A occurs if and only if one of the outcomes in $A \cap B$ occurs.

Independent Events

Let A and B be events of some sample space Ω.

- We say A and B are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$.
- We say A and B are dependent if $\Pr(A \cap B) \neq \Pr(A)\Pr(B)$.

Note: Disjoint events are not necessarily independent!

Independence versus conditional probabilities:
Suppose A and B are independent events.

- If $\Pr(B) > 0$, then $\Pr(A|B) = \Pr(A)$.
- If $\Pr(A) > 0$, then $\Pr(B|A) = \Pr(B)$.

Independence of multiple events
Events $A_1, A_2, \ldots, A_n$ are mutually independent if for every subset of indices $\{i_1, i_2, \ldots, i_k\}$ (for $k = 2, 3, \ldots, n$),

$$\Pr(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \Pr(A_{i_1})\Pr(A_{i_2}) \cdots \Pr(A_{i_k}).$$

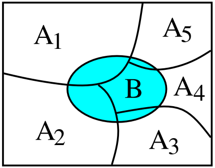
Law of total probability and Bayes theorem

Let $A_1, \ldots, A_k$ be **mutually exclusive** and **exhaustive** events in some sample space Ω.

- $A_1, \ldots, A_k$ are exhaustive if $A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$.
- $A_1, \ldots, A_k$ are mutually exclusive if $A_i \cap A_j = \varnothing$ for all $i \neq j$.

The **law of total probability** states that for any event B,

$$\Pr(B) = \sum_{i=1}^{k} \Pr(B|A_i) P(A_i)$$



Bayes' theorem states that if B is an event such that $\Pr(B) > 0$, then for every $j = 1, \ldots, k$,

$$\Pr(A_j|B) = \frac{\Pr(B|A_j)\Pr(A_j)}{\Pr(B)} = \frac{\Pr(B|A_j)\Pr(A_j)}{\sum_{i=1}^{k} \Pr(B|A_i)\Pr(A_i)}$$

Random variables

Recall: A random variable is a real-valued function on a sample space Ω of some experiment.

A random variable X is called discrete if X can take only a finite number k of different values $x_1, \ldots, x_k$, or, at most, an infinite sequence of different values $x_1, x_2, x_3, \ldots$

A random variable X is called continuous if the following two conditions hold:

- The set of all possible values for X is either a single interval on the real line (possibly the entire real line), or a union of disjoint intervals on the real line.
- $\Pr(X = x) = 0$ for every possible value x.

Useful Fact: Any real-valued function of a R.V. is a R.V.!

- More generally, any real-valued function of any number of R.V.'s on the same sample space is a R.V.!

Probability distributions

The probability distribution of a R.V. X is the collection of all probabilities of the form $\Pr(X \in C)$, for all sets C of real numbers. (This definition is for *any* R.V., not just discrete or continuous R.V.'s.)

Interpretation: For any set C of real numbers, the probability distribution of X gives the probability $\Pr(X \in C)$ of how likely the random variable X takes on values in C.

There are other ways to represent the same information given by the probability distribution of a random variable:

- probability mass function (only for discrete R.V.)
- probability density function (only for continuous R.V.)
- cumulative distribution function (for *any* R.V.)

The pmf/pdf/cdf of a R.V.

Let X be a random variable.

- If X is **discrete**, then the pmf of X is $p(x) = \Pr(X = x)$.
- If X is **continuous**, then the pdf of X is a function $f(x)$, such that for any two numbers a and b with $a \leq b$,

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\, dx$$

- For **arbitrary** R.V. X, the cdf of X is $F(x) = \Pr(X \leq x)$.
  - If X is a **discrete** R.V. with **pmf** $p(x)$, then its cdf is

$$F(x) = \Pr(X \leq x) = \sum_{y: y \leq x} p(y).$$

  - If X is a **continuous** R.V. with **pdf** $f(x)$, then its cdf is

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(u)\, du.$$

Some useful properties of pmf/pdf/cdf

Let X be a random variable.

- If $F(x)$ is the **cdf** of X, then the following hold:
  - $\Pr(X > a) = 1 - F(a)$ for all $a \in \mathbb{R}$.
  - If $x_1 < x_2$, then $F(x_1) \leq F(x_2)$ (i.e. $F(x)$ is non-decreasing).
  - $\Pr(a < X \leq b) = F(b) - F(a)$ whenever $a < b$.
  - $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

- If X is discrete, then its **pmf** $p(x)$ satisfies:
  - $p(x) \geq 0$ for all x.
  - $\sum_{x \in D} p(x) = 1$, where D is the set of all possible values.

- If X is continuous, then its **pdf** $f(x)$ satisfies:
  - $f(x) \geq 0$ for all x. (Density cannot be negative.)
  - $\int_{-\infty}^{\infty} f(x)\, dx = 1$. (Area under the curve of $f(x)$ is 1.)
  - **Important Fact:** If $F(x)$ is the cdf whose derivative $F'(x)$ exists at $x = x_0$, then $f(x_0) = F'(x_0)$.

Joint distributions

The joint distribution of **any** R.V.'s X and Y is the collection of all probabilities of the form $\Pr\left((X,Y) \in C\right)$, for all sets $C \subseteq \mathbb{R}^2$.

Ways to describe the joint distribution of X and Y:

- joint probability mass function (only for discrete R.V.'s)

$$p(x,y) = \Pr(X = x \text{ and } Y = y) = \Pr\left((X,Y) = (x,y)\right).$$

- joint probability density function (only for continuous R.V.'s)

$$\Pr\left((X,Y) \in A\right) = \iint_A f(x,y)\,dx\,dy, \qquad \text{where } A \subseteq \mathbb{R}^2.$$

- joint cumulative distribution function (for *any* R.V.'s)

$$F(x,y) = \Pr(X \le x, Y \le y), \qquad \text{for } -\infty < x, y < \infty.$$

  - $F(a,b) = \sum_{x \le a} \sum_{y \le b} p(x,y)$. (discrete R.V.'s case)
  - $F(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\,dx\,dy$. (continuous R.V.'s case)

Some useful properties of joint pmf/pdf/cdf

Let X and Y be random variables.

- If $F(x, y)$ is the **joint cdf** of X, Y, then the following hold:
  - If $x_1 < x_2$, then $F(x_1, y) \leq F(x_2, y)$ for any $y \in \mathbb{R}$.
  - If $y_1 < y_2$, then $F(x, y_1) \leq F(x, y_2)$ for any $x \in \mathbb{R}$.
  - If $a < b$ and $c < d$, then $\Pr(a < X \leq b, c < Y \leq d)$ equals

$$F(b, d) - F(a, d) - F(b, c) + F(a, c).$$

- If X, Y are discrete, then its **joint pmf** $p(x, y)$ satisfies:
  - $p(x, y) \geq 0$ for all x, y.
  - $\sum_{x \in D_X} \sum_{y \in D_Y} p(x, y) = 1$.
- If X, Y are continuous, then its **joint pdf** $f(x, y)$ satisfies:
  - $f(x, y) \geq 0$ for all x, y.
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$.
  - **Important Fact:** If $F(x, y)$ is the joint cdf such that $\frac{\partial^2 F}{\partial x \partial y}(x, y)$ exists at $(x, y) = (x_0, y_0)$, then $f(x_0, y_0) = \frac{\partial^2 F}{\partial x \partial y}(x_0, y_0)$.

Marginal pmf/pdf/cdf

If X and Y are **discrete** R.V.'s with joint pmf $p(x, y)$, then:

- The marginal pmf of X is $p_X(x) = \sum_{y \in D_Y} p(x, y)$;
- The marginal pmf of Y is $p_Y(x) = \sum_{x \in D_X} p(x, y)$;

where D_X and D_Y are the sets of possible values for X and Y.

If X and Y are **continuous** R.V.'s with joint pdf $f(x, y)$, then:

- The marginal pdf of X is $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$.
- The marginal pdf of Y is $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$.

If X and Y are **arbitrary** R.V.'s with joint cdf $F(x, y)$, then:

- The marginal cdf of X is $F_X(x) = \lim_{y \to \infty} F(x, y)$.
- The marginal cdf of Y is $F_Y(y) = \lim_{x \to \infty} F(x, y)$.

Conditional distribution/pmf/pdf

Let $C' \subseteq \mathbb{R}$, and let X and Y be **arbitrary** R.V.'s. The conditional distribution of X given $Y \in C'$ is the collection of all conditional probabilities of the form $\Pr(X \in C | Y \in C')$ for all sets $C \subseteq \mathbb{R}$.

If X and Y are **discrete** R.V.'s with joint pmf $p(x, y)$, and if $y \in \mathbb{R}$ such that $p_Y(y) > 0$, then the conditional pmf of X given $Y = y$, is the function $p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)}$.

- For fixed y, $\sum_{x \in D_X} p_{X|Y}(x|y) = 1$. (A conditional pmf is a pmf!)

If X and Y are **continuous** R.V.'s with joint pdf $f(x, y)$, and if $y \in \mathbb{R}$ such that $f_Y(y) > 0$, then the conditional pdf of X given $Y = y$, is the function $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$.

- For fixed y, $\int_{-\infty}^{\infty} f_{X|Y}(x|y)\, dx = 1$. (A conditional pdf is a pdf!)

Law of total probability for two R.V.'s

The **law of total probability for discrete R.V.'s** states that for two **discrete** R.V.'s X and Y,

$$p_X(x) = \sum_{y \in D_Y} p_{X|Y}(x|y) p_Y(y),$$

where D_Y is the set of possible values for Y.

- If we know the marginal pmf of Y, and the conditional pmf of X given $Y = y$, then we can find the marginal pmf of X.

The **law of total probability for continuous R.V.'s** states that for two **continuous** R.V.'s X and Y,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\, dy.$$

- If we know the marginal pdf of Y, and the conditional pdf of X given $Y = y$, then we can find the marginal pdf of X.

Bayes' theorem for two R.V.'s

The **Bayes' theorem for discrete R.V.'s** states that for two **discrete** R.V.'s X and Y,

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}.$$

- Theorem relates the two conditional pmf's ($Y|X$ and $X|Y$).

The **Bayes' theorem for continuous R.V.'s** states that for two **continuous** R.V.'s X and Y,

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}.$$

- Theorem relates the two conditional pdf's ($Y|X$ and $X|Y$).

Expectation of a R.V.

Let X be a random variable, and let $h : \mathbb{R} \to \mathbb{R}$ be any function.

- If X is **discrete** with pmf $p(x)$, and has the set of possible values D, then the expectation of X (if it exists) is

$$\mathbf{E}[X] = \sum_{x \in D} x \cdot p(x).$$

  - More generally, $\mathbf{E}[h(X)] = \sum_{x \in D} h(x) \cdot p(x)$ (if it exists).

- If X is **continuous** with pdf $f(x)$, then the expectation of X (if it exists) is

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

  - More generally, $\mathbf{E}[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx$ (if it exists).

(We use "expectation", "expected value" and "mean" interchangeably.)

Expectation $\mathbf{E}[XY]$

Let X and Y be R.V.'s, and let $h(x, y)$ be a bivariate function.

- If X, Y are **discrete** with joint pmf $p(x, y)$, and have the sets of possible values D_X, D_Y respectively, then

$$\mathbf{E}[XY] = \sum_{x \in D_X} \sum_{y \in D_Y} xy \cdot p(x, y) \quad \text{(if it exists).}$$

  - More generally, $\mathbf{E}[h(X, Y)] = \sum_{x \in D_X} \sum_{y \in D_Y} h(x, y) \cdot p(x, y)$ (if it exists).

- If X, Y are **continuous** with joint pdf $f(x, y)$, then

$$\mathbf{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f(x, y) \, dx \, dy.$$

  - More generally, $\mathbf{E}[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y) \, dx \, dy$ (if it exists).

Some useful properties of expectation

Let X, Y be R.V.'s with finite means.

- $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$ for any finite constants a, b.
- If X and Y are **independent**, then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.
  - **Note:** In general, $\mathbf{E}[X^2] \neq \mathbf{E}[X]\mathbf{E}[X]$.

Let $X_1, \dots, X_n$ be R.V.'s with finite means.

- If $a_1, \dots, a_n, b$ are finite constants, then

  $$\mathbf{E}[a_1X_1 + \cdots + a_nX_n + b] = a_1\mathbf{E}[X_1] + \cdots + a_n\mathbf{E}[X_n] + b.$$

  - In particular, "mean of sum" = "sum of means":

    $$\mathbf{E}[X_1 + \cdots + X_n] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n].$$

- If $X_1, \dots, X_n$ are **independent**, then

  $$\mathbf{E}[X_1 \cdots X_n] = \mathbf{E}[X_1] \cdots \mathbf{E}[X_n].$$

Conditional expectation

Idea: Conditional distributions have conditional expectations.

Let X, Y be R.V.'s, and let $C' \subseteq \mathbb{R}$ such that $\Pr(Y \in C') > 0$.
Definition: The conditional expectation of X given $Y \in C'$ is:

$$\mathbf{E}[X|Y \in C'] = \mathbf{E}\left[\begin{array}{c}\text{conditional distribution}\\ \text{of } X \text{ given } Y \in C'.\end{array}\right]$$

- We can think of $\mathbf{E}[X|Y \in C']$ as a function in terms of C'.
  - Different values of C' give different values for $\mathbf{E}[X|Y \in C']$.
- Similarly, we can think of $\mathbf{E}[X|Y = y]$ as a function of y.
  - Different values of y give different values for $\mathbf{E}[X|Y = y]$.

More generally, we can think of $\mathbf{E}[X|Y]$ as a **function of** Y.

- In other words, $\mathbf{E}[X|Y]$ is a random variable!

The **law of total probability for expectations** states that if X and Y are **arbitrary** R.V.'s such that X has finite mean, then

$$\mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[X].$$

Variance of a R.V.

Definition: Let X be an **arbitrary** R.V. with finite mean μ_X. Then the variance of X, if it exists, is

$$\mathrm{var}(X) = \mathbf{E}[(X - \mu_X)^2],$$

and the standard deviation of X is $\sqrt{\mathrm{var}(X)}$.

- μ_X must exist and be finite, for $\mathrm{var}(X)$ to make sense.
- If $\mu_X = \pm\infty$ or does not exist, then $\mathrm{var}(X)$ does not exist.

Very useful formula: $\mathrm{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

Some properties of variance:

- $\mathrm{var}(aX + b) = a^2\mathrm{var}(X)$ for any finite constants a, b.
- If $X_1, \ldots, X_n$ are **independent** R.V.'s with finite means, then $\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)$.

Covariance and Correlation

Let X and Y be R.V.'s with finite means μ_X and μ_Y respectively. The covariance of X and Y is $\mathrm{cov}(X, Y) = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)]$, provided that this expectation $\mathbf{E}[(X - \mu_X)(Y - \mu_Y)]$ exists.

- $\mathrm{cov}(X, Y) =$ "how strongly X and Y are **linearly** related".

Very useful formula: $\mathrm{cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$.

- **Note:** $\mathrm{var}(X) = \mathrm{cov}(X, X)$.

If X and Y have finite variances σ_X^2 and σ_Y^2 respectively, then the correlation of X and Y is

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

- **Fact:** $-1 \leq \rho(X, Y) \leq 1$ (by Cauchy–Schwarz inequality).
- Correlation does not depend on scale or choice of units.

Independence of R.V.'s

A collection $\{X_1, \ldots, X_n\}$ of R.V.'s is called independent if $\{X_1 \in C_1\}, \ldots, \{X_n \in C_n\}$ are **mutually independent events** for all possible sets $C_1, \ldots, C_n \subseteq \mathbb{R}$.

Theorem: (Useful criteria for independence of R.V.'s)

- A collection of discrete R.V.'s is independent if and only if the joint pmf is the **product of the marginal pmf's**.
  - i.e. $p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$ for all $x_1, \ldots, x_n \in \mathbb{R}$.
- A collection of continuous R.V.'s is independent if and only if the joint pdf is the **product of the marginal pdf's**.
  - i.e. $f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ for all $x_1, \ldots, x_n \in \mathbb{R}$.
- A collection of arbitrary R.V.'s is independent if and only if the joint cdf is the **product of the marginal cdf's**.
  - i.e. $F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$ for all $x_1, \ldots, x_n \in \mathbb{R}$.

Useful Fact: If X and Y are any R.V.'s with $\text{cov}(X, Y) \neq 0$, then X and Y are **not** independent.

- Warning: zero covariance does not imply independence!

Moments and moment generating functions

Let X be a random variable, and let k be any positive integer.

- $\mathbf{E}[X^k]$ is called the *k-th moment* of X.
- $\mathbf{E}[(X-\mu)^k]$ is called the *k-th central moment* of X.
  - For this to make sense, $\mu = \mathbf{E}[X]$ should be finite.
- $\mathbf{E}[X]$ = first moment; $\mathrm{var}(X)$ = second central moment.

The moment generating function (mgf) of X is $\psi(t) = \mathbf{E}[e^{tX}]$.

- $\psi(t)$ (if it exists) depends only on the distribution of X.
- **Technical Note:** The domain of $\psi(t)$ is the set of all real values of t such that $\mathbf{E}[e^{tX}]$ exists.

Useful Fact: If X and Y are R.V.'s with the same distribution, then X and Y must have the same mgf (if it exists).

- In other words, to check if X and Y are identically distributed, it suffices to check if their mgf's coincide (if they exist).

Special distributions

Bernoulli distribution (with parameter p)

- **Main Use:** To model a **single Bernoulli trial**.
  - e.g. a single coin toss: 1 (heads) and 0 (tails).
  - The parameter p is usually called the success rate.

Binomial distribution (with parameters n and p)

- **Main Use:** To model the **sum of n Bernoulli trials**.
  - e.g. the number of heads in n coin tosses.

Geometric distribution (with parameter p)

- **Main Use:** To model the **number of failed trials** in a Bernoulli process immediately before the first success.
  - e.g. the number of coin tosses before getting the first heads.
  - Every geometric R.V. X has the memoryless property, i.e. $\Pr(X = k + t | X \geq k) = \Pr(X = t)$ for all integers $k, t \geq 0$.

Poisson distribution (with parameter λ)

- **Common Use:** To model the **number of occurrences** of an event during a **fixed time period**.
  - e.g. number of people visiting Apple Store in the past hour.

Special distributions (continued)

Exponential distribution (with parameter λ)

- **Common Use:** To model the **elapsed time** between two **successive events**.
  - e.g. time between two people visiting Apple Store.
  - Every exponential R.V. X has the memoryless property, i.e. $\Pr(X \geq t + h | X \geq t) = \Pr(X \geq h)$ for all $t > 0$, $h > 0$.

Normal distribution (with parameters μ and σ)

- **Main Use:** To model symmetric "bell-shaped" distributions.
  - e.g. heights of individuals in a population.
  - $X \sim N(\mu, \sigma^2)$ means X is normally distributed with mean μ and variance σ^2.
  - Any non-zero linear function of a normal R.V. is a normal R.V.

Bivariate normal distribution (with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$)

- **Main Use:** To model correlated normal distributions.
  - e.g. heights and weights of individuals in a population.
  - The parameter ρ represents the **correlation** of the two R.V.'s.
  - **Theorem:** If X and Y have a bivariate normal distribution, then $\mathrm{cov}(X, Y) = 0$ if and only if X, Y are independent.

Special Case: Standard normal distribution

The standard normal distribution is the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

- A standard normal R.V. is usually denoted by Z.
  - We write $Z \sim N(0, 1)$ to mean Z is a standard normal R.V.
- The **cdf** of Z is usually written as $\Phi(z)$.
  - $\Phi(z)$ is usually called the standard normal cdf.
- Some useful properties:
  - The graph of $\phi(z)$ is symmetric about its mean $z = 0$.
  - $\Phi(-z) = 1 - \Phi(z)$ for all real numbers z.
  - $\Phi^{-1}(p) = -\Phi^{-1}(1 - p)$ for all real numbers $0 < p < 1$.

Very Important Fact: If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ has the standard normal distribution.

- If $X \sim N(\mu, \sigma^2)$ with cdf $F(x)$, then $F(x) = \Phi(\frac{x-\mu}{\sigma})$.
- To do computations involving normal R.V.'s, we transform normal R.V.'s into standard normal R.V.'s.

Markov's inequality and Chebyshev's inequality

Theorem: (**Markov's inequality**) Let X be any R.V. satisfying $\Pr(X \geq 0) = 1$. If the expectation $\mathbf{E}[X]$ exists, then for every real number $t > 0$,

$$\Pr(X \geq t) \leq \frac{\mathbf{E}[X]}{t}.$$

- **Intuition:** The event $\{X \geq t\}$ should have a small probability when t is very large. Markov's inequality gives a bound for how small this probability could be.

Theorem: (**Chebyshev's inequality**) Let X be any R.V. with a finite mean. If its variance $\mathrm{var}(X)$ exists, then for every real $t > 0$,

$$\Pr(|X - \mathbf{E}[X]| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}.$$

- **Intuition:** The probability that X deviates a lot from the mean should be small. Chebyshev's inequality gives a bound for how small this probability could be.

Sample Mean

Definition: Let $X_1, \ldots, X_n$ be R.V.'s. The sample mean of $X_1, \ldots, X_n$ is the R.V. $\frac{X_1 + \cdots + X_n}{n}$, i.e. the mean of $\{X_1, \ldots, X_n\}$.

- The sample mean is commonly denoted by $\overline{X}_n$.
- If $X_1, \ldots, X_n$ are n **independent** R.V.'s, such that each X_i has the same distribution, then we say that $X_1, \ldots, X_n$ are independent and identically distributed or iid.
- A random sample is a collection $\{X_1, \ldots, X_n\}$ of **iid** R.V.'s, and the number n is called the sample size.

Very Important Consequence of Chebyshev's inequality: Suppose $\overline{X}_n$ is the sample mean of n iid R.V.'s $X_1, \ldots, X_n$, each with mean μ and variance σ^2.

- **Fact:** $\mathbf{E}[\overline{X}_n] = \mu$ and $\mathrm{var}(\overline{X}_n) = \frac{\sigma^2}{n}$.
- Chebyshev's inequality says that $\Pr(|\overline{X}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2}$.
- In particular, for fixed $\varepsilon > 0$, as the sample size $n \to \infty$, the probability $\Pr(|\overline{X}_n - \mu| \geq \varepsilon)$ approaches 0.

Different kinds of convergences

Let $X_1, X_2, X_3, \ldots$ be an infinite sequence of R.V.'s.

- We say that $X_1, X_2, X_3, \ldots$ converges in probability to some R.V. X, which we write as $X_n \xrightarrow{p} X$, if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| < \varepsilon) = 1.$$

  - A real number r can be treated as a discrete R.V. with exactly one possible value r. In particular, $\mathbf{E}[r] = r$, and $\mathrm{var}(r) = 0$.

- We say that $X_1, X_2, X_3, \ldots$ converges almost surely to X if

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

  - i.e. $\Pr(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$.
  - We usually write X_n converges a.s. to X, or write $X_n \xrightarrow{\text{a.s.}} X$.

- Let $F_i(x)$ be the cdf of each X_i, and let $F(x)$ be the cdf of X. We say that $X_1, X_2, X_3, \ldots$ converges in distribution to X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

  for all $x \in \mathbb{R}$ at which F is continuous. We write $X_n \xrightarrow{d} X$.
  - The distribution of X is called the asymptotic distribution.

Law of large numbers, Central limit theorem

Let $X_1, X_2, X_3, \ldots$ be an infinite sequence of **iid** R.V.'s.

- Suppose each X_i has finite mean μ and finite variance σ^2.
- For every n, let $\overline{X}_n$ be the sample mean of $\{X_1, \ldots, X_n\}$.

Fact: $X_n \xrightarrow{\text{a.s.}} X$ implies $X_n \xrightarrow{\text{p}} X$ implies $X_n \xrightarrow{\text{d}} X$.

Theorem: (**Weak law of large numbers**) $\overline{X}_n \xrightarrow{\text{p}} \mu$.

Theorem: (**Strong law of large numbers**) $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$.

Theorem: (**Central limit theorem**) Let $Z \sim N(0,1)$, and define $Z_1, Z_2, Z_3, \ldots$ by $Z_n = \frac{(\overline{X}_n - \mu)\sqrt{n}}{\sigma}$ for each n. Then $Z_n \xrightarrow{\text{d}} Z$.

- In other words, the **asymptotic distribution** of the sequence $Z_1, Z_2, Z_3, \ldots$ is the standard normal distribution.
- **Interpretation:** For large n, $\overline{X}_n$ is approximately normal.

Summary

- Conditional probability and independent events
- Law of total probability, Bayes' theorem
- Probability/joint/conditional distributions
- Joint/marginal pmf/pdf/cdf
- Expectation and variance
- Covariance, correlation, independence
- Moments, moment generating functions
- Special distributions (Poisson, exponential, etc.)
- Markov's inequality, Chebyshev's inequality
- Law of large numbers, central limit theorem

All the best for your mid-term exam tomorrow!
(2–4pm MPH, be at least 10 minutes early!)