Course Overview

Three sections:

- Monday 2:30pm-3:45pm Wednesday 1:00pm-2:15pm Thursday 2:30pm-3:45pm

Goal:

- Review some basic (or not so basic) concepts in probability and statistics
- Good preparation for CME 308

Syllabus:

- Basic probability, including random variables, conditional distribution, moments, concentration inequality
- Convergence concepts, including three types of convergence, WLLN, CLT, delta method
- Statistical inference, including fundamental concepts in inference, point estimation, MLE

Section 1: Basic probability theory

Dangna Li

ICME

September 14, 2015

Outline

Random variables: discrete and continuous

- For our purposes, random variables will be one of two types: discrete or continuous.
- A random variable X is discrete if its set of possible values $\mathbb{X}$ is finite or countably infinite.
- A random variable X is continuous if its possible values form an uncountable set (e.g., some interval on $\mathbb{R}$) and the probability that X equals any such value exactly is zero.
- Examples:
  - Discrete: binomial, geometric, Poisson, and discrete uniform random variables
  - Continuous: normal, exponential, beta, gamma, chi-squared, Student's t, and continuous uniform random variables

The probability density (mass) function

- *pmf*: The probability mass function (pmf) of a discrete random variable X is a nonnegative function $f(x) = P(X = x)$, where x denotes each possible value that X can take. It is always true that $\sum_{x \in \mathbb{X}} f(x) = 1$.
- *pdf*: The probability density function (pdf) of a continuous random variable X is a nonnegative function $f(x)$ such that $\int_a^b f(x)dx = P(a \leq \text{X} \leq b)$ for any $a, b \in \mathbb{R}$. It is always true that $\int_{-\infty}^{\infty} f(x)dx = 1$.

The cumulative distribution function

The cumulative distribution function (cdf) of a random variable X is $F(x) = P(X \leq x)$.

- If X is discrete, then $F(x) = \sum_{t \in \mathbb{X}: t \leq x} f(t)$, and so the cdf consists of constant sections separated by jump discontinuities.
- If X is continuous, then $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$, and so the cdf is a continuous function regardless of the continuity of f.

Note

The cdf is a more general description of a random variable than the pmf or pdf, since it has a single definition that applies for both discrete and continuous random variables.

A common mistake in probability

A random variable is not the same thing as its distribution.

One might find the following helpful in distinguishing these two concepts

- A distribution can be thought of as a blueprint for generating r.v.s. Confusing a distribution with that r.v. is like confusing a blueprint of a house with the house itself. *The word is not the thing, the map is not the territory.*
- It is possible to have two r.v.s which have the same distribution but never equal to each other.

Conditional probability

The conditional probability of event A given event B is defined as

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

Quiz

Is it true that $P(A|B)$ always larger than $P(A)$? or less?

Conditional distribution

If X and Y are both discrete random variables with joint probability mass function $p_{X,Y}(x, y)$, then the conditional probability mass function of X given Y is given by:

$$P(X = x|Y = y) = p_{X|Y}(x|y) := \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

If X and Y are both continuous random variables with joint density function $f_{X,Y}(x, y)$, the conditional probability density function of X given Y is given by:

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

Another common mistake

A Conditional p.d.f. is not the result of conditioning on a set of probability zero.

- The conditional p.d.f. $f_{X|Y}(x|y)$ of X given $Y = y$ is the p.d.f. we would use for X if we were to learn that $Y = y$. So that $\int_A f_{X|Y}(x|y) = P(X \in A|Y = y)$ for any set $A \in \mathbb{R}$.
- This sounds as if we were conditioning on the event $Y = y$, which has zero probability if Y has a continuous distribution.
- However, this is not technically correct. $P(X \in A|Y = y)$ can not even be properly defined using our definition of conditional probability.
- Actually, the value of $f_{X|Y}(x|y)$ is a limit:

$$f_{X|Y}(x|y) = \lim_{\epsilon \to 0} \frac{\partial}{\partial x} P(X \leq x|y - \epsilon < Y < y + \epsilon)$$

Outline

Expectation

The expectation $\mathbb{E}[X]$ of a continuous random variable X is defined as:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) dx$$

Similarly, the expectation of a function $g(\cdot)$ of X can be computed as (LOTUS):

$$\mathbb{E}[g(X)] = \int_{R} g(x) f(x) dx$$

Quiz

Does $\mathbb{E}[X]$ always exist?

Variance

The variance $\mathrm{Var}[X]$ of a random variable X is defined as

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

An equivalent (and typically easier) formula is

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

Similarly, the variance of a function $g(X)$ of a random variable X is

$$\mathrm{Var}[g(X)] = \mathbb{E}\left[g(X)^2\right] - (\mathbb{E}[g(X)])^2$$

Quiz

If you want to implement $\mathrm{Var}[X]$ on a computer, which formula would you chose?

Covariance

The covariance $\mathrm{Cov}\left[X, Y\right]$ of a random variable X and a random variable Y is defined as

$$\mathrm{Cov}\left[X, Y\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right].$$

An equivalent (and typically easier) formula is

$$\mathrm{Cov}\left[X, Y\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

Similarly, the covariance of $g(X)$ and $h(Y)$ is

$$\mathrm{Cov}\left[g(X), g(Y)\right] = \mathbb{E}\left[g(X)h(Y)\right] - \mathbb{E}\left[g(X)\right]\mathbb{E}\left[h(Y)\right]$$

Outline

Important properties of expectation

- *Linearity*:
  $\mathbb{E}[a + bg(X) + ch(Y)] = a + b\mathbb{E}[g(X)] + c\mathbb{E}[h(Y)]$
  In particular, for a sequence of random variables $\{X_i\}_{i=1}^n$,
  $$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$
- *The fundamental bridge*:
  Let $\mathbb{I}(\cdot)$ be the indicator function for some random event A, then
  $$\mathbb{E}[\mathbb{I}(A)] = P(A)$$

Linearity of expectation: An example

A group of n people play "Secret Santa" as follows: each puts his or her name on a slip of paper in a hat, picks a name randomly from the hat (without replacement), and then buys a gift for that person. Unfortunately, they overlook the possibility of drawing one's own name, so some may have to buy gifts for themselves. Assume $n \geq 2$.

Find the expected number of pairs of people, A and B, such that A picks $B's$ name and B picks $A's$ name (where $A \neq B$ and order doesn't matter).

Important properties of variance and covariance

- $\mathrm{Var}\left[a + bg(X)\right] = b^2 \mathrm{Var}\left[g(X)\right]$
- $\mathrm{Cov}\left[a + bg(X), h(Y)\right] = b\mathrm{Cov}\left[g(X), h(Y)\right]$
- If X and Y are independent, then

$$\mathrm{Cov}\left[g(X), h(Y)\right] = 0$$

- If X and Y are independent, then

$$\mathrm{Var}\left[g(X) + h(Y)\right] = \mathrm{Var}\left[g(X)\right] + \mathrm{Var}\left[h(Y)\right]$$

Outline

Conditional expectation

The conditional expectation of a continuous random variable X given another random variable Y is defined as:

$$\mathbb{E}\left[X|Y=y\right] = \int_{\mathbb{R}} x f_{X|Y}(x|y)dx$$

where $f_{X|Y}(\cdot|\cdot) = \frac{f_{X,Y}(\cdot,\cdot)}{f_Y(\cdot)}$ is the conditional probability density function of X given Y.

Remarks

- Notice that computing $\mathbb{E}\left[X|Y=y\right]$ yields (in general) different results for different values of y. Thus, $\mathbb{E}\left[X|Y=y\right]$ is a function of y (and not a random variable).
- If we plug the random variable Y into this function, which does yield a random variable. This random variable is what we mean when we write $\mathbb{E}\left[X|Y\right]$.

Conditional variance

The conditional variance of a continuous random variable X given another random variable Y is defined as:

$$\mathrm{Var}\left[X|Y=y\right] = \mathbb{E}\left[X^2|Y=y\right] - \left(\mathbb{E}\left[X|Y=y\right]\right)^2$$

Remarks

- Again, we might consider either $\mathrm{Var}\left[X|Y=y\right]$, which is a function of y, or $\mathrm{Var}\left[X|Y\right]$ which is a random variable.

Outline

Important properties of conditional expectation

- *Linearity*: $\mathbb{E}[X_1 + X_2|Y] = \mathbb{E}[X_2|Y] + \mathbb{E}[X_2|Y]$
- *Independence*: if X and Y are independent: $\mathbb{E}[X|Y] = \mathbb{E}[X]$
- *Taking out what's known*:

$$\mathbb{E}[h(Y)X|Y] = h(Y)\mathbb{E}[X|Y]$$

- *Law of Total Expectation*:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

- *Law of Total Variance*:

$$\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathbb{E}[X|Y]]$$

Outline

Concentration inequality

- Concentration inequalities provide probability bounds on how a random variable deviates from some value (e.g., its expectation).
- Most concentration inequalities are about the concentrating behavior of the sum of a sequence of *iid* random variables. But such behavior is shared by other functions of independent random variables as well.
- As an example, the laws of large numbers(which we will see in the next section) states that sums of independent random variables are, under very mild conditions, close to their expectation with a large probability.

Markov inequality

Let X be any *nonnegative* integrable random variable then for all $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Example: application of Markov inequality

Consider a biased coin, which lands heads with probability $1/10$. Suppose the coin is flipped 200 times consecutively. Give an upper bound on the probability that it lands heads at least 120 times.

Solution:

The total number of heads is a binomial random variable X, with parameters $p = 1/10$ and $n = 200$. Thus, the expected number of heads is

$$\mathbb{E}[X] = np = 20$$

By Markov inequality, the probability of at least 120 heads is

$$P(X \geq 120) \leq \frac{\mathbb{E}[X]}{120} = \frac{20}{120} = 1/6$$

Outline

Chebyshev inequality

Let X be any integrable random variable then for all $a > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$$

Outline

Chernoff bound

Chernoff bound

Suppose we conduct a sequence of n *iid* Bernoulli trials, with probability p of landing head. Let X be the total number of heads in n trials. Then $X \sim \text{Bin}(n, p)$ (recall that $\mathbb{E}[X] = np$). Then

$$P(X \geq (1+\delta)\mathbb{E}[X]) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}$$

If we let $\delta = 5$, one can show that :

$$P(X \geq 6\mathbb{E}[X]) \leq 2^{-(6\mathbb{E}[X])}$$

Example: application of Chernoff bound

If we apply the Chernoff bound on the previous example, we get:

$$P(X \geq 120) = P(X \geq 6\mathbb{E}[X]) \geq 2^{-6E(X)} = 2^{-(6\times 20)} = 2^{-120}$$

which is vastly better than the one obtained from Markov inequality.