Data Driven Sampling Methods

Lecture 2: An overview of probability

Hongwei YUAN

University of Macau

Definition (Probability)

Suppose S is a sample space associated with an experiment. To every event A in S (A is a subset of S), we assign a number, $P(A)$, called the probability of A, so that the following axioms hold:

1 : $P(A) \geq 0$.

2 : $P(S) = 1$.

3 : If $A_1, A_2, A_3, \ldots$ form a sequence of pairwise mutually exclusive events in S (that is, $A_i \cap A_j = \emptyset$ if $i \neq j$), then

$$P(A_1 \cup A_2 \cup \ldots \cup \ldots) = P(A_1) + P(A_2) + \ldots.$$

Definition (Conditional probability)

The conditional probability of an event A, given that an event B has occurred, is equal to

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

provided $P(B) > 0$. [The symbol $P(A|B)$ is read 'probability of A given B.']

Definition (Independence)

Two events A and B are said to be independent if any one of the following holds:

$$P(A \cap B) = P(A)P(B).$$

Otherwise, the events are said to be dependent.

Theorem (The Multiplicative Law of Probability)

The probability of the intersection of two events A and B is

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

If A and B are independent, then

$$P(A \cap B) = P(A)P(B).$$

Theorem (The Additive Law of Probability)

The probability of the union of two events A and B is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

If A and B are mutually exclusive events, $P(A \cap B) = 0$ and

$$P(A \cup B) = P(A) + P(B).$$

Definition (Partition)

For some positive integer k, let the sets $B_1, B_2, ..., B_k$ be such that 1. $S = B_1 \cup B_2 \cup ... \cup B_k$. 2. $B_i \cap B_j = \emptyset$, for $i \neq j$. Then the collection of sets $\{B_1, B_2, ..., B_k\}$ is said to be a partition of S.

Definition (Total probability law)

Assume that $\{B_1, B_2, ..., B_k\}$ is a partition of S such that $P(B_i) > 0$, for i = 1, 2,..., k. Then for any event A

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$

Definition (Bayes' Rule)

Assume that $\{B_1, B_2, ..., B_k\}$ is a partition of S such that $P(B_i) > 0$, for i = 1, 2,..., k. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

Definition (Discrete random variable)

A random variable Y is said to be discrete if it can assume only a finite or countably infinite number of distinct values.

Definition (Probability distribution)

The probability distribution for a discrete variable Y can be represented by a formula, a table, or a graph that provides

$$p(y) = P(Y = y) \text{ for all } y.$$

Theorem

For any discrete probability distribution, the following must be true: (1) $0 \leq p(y) \leq 1$ *for all y. (2)* $\sum_y p(y) = 1$*, where the summation is over all values of y with nonzero probability.*

Definition (Expected value (Expectation))

Let Y be a discrete random variable with the probability function p(y). Then the expected value of Y , E(Y), is defined to be

$$E(Y) = \sum_y yp(y).$$

Theorem

Let Y be a discrete random variable with probability function $p(y)$ and $g(Y)$ be a real-valued function of Y. Then the expected value of $g(Y)$ is given by

$$E[g(Y)] = \sum_y g(y)p(y).$$

Definition (Variance)

If Y is a random variable with mean E(Y) $= \mu$, the variance of a random variable Y is defined to be the expected value of $(Y - \mu)^2$. That is,

$$V(Y) = E[(Y - \mu)^2].$$

The standard deviation of Y is the positive square root of V(Y).

Linear property of Expectation

Theorem

Let Y be a discrete random variable with probability function $p(y)$, $g(Y)$ be a function of Y, and c be a constant. Then

$$E[cg(Y)] = cE[g(Y)]$$

.

Theorem

Let Y be a discrete random variable with probability function $p(y)$ and $g_1(Y)$, $g_2(Y)$, ... , $g_k(Y)$ be k functions of Y . Then

$$E[g_1(Y)+g_2(Y)+...+g_k(Y)] = E[g_1(Y)]+E[g_2(Y)]+...+E[g_k(Y)].$$

Theorem

Let Y be a discrete random variable with probability function $p(y)$ and mean $E(Y) = \mu$; then

$$V(Y) = \sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2.$$

The probability that y successes occurs in n trials:

Definition (Binomial distribution)

A random variable Y is said to have a binomial distribution based on n trials with success probability θ if and only if

$$p(y) = C_y^n \theta^y (1-\theta)^{n-y}; \qquad y = 0, 1, 2, ..., n, 0 \leq \theta \leq 1,$$

where $C_y^n = \frac{n!}{y!(n-y)!}$.

Theorem

Let Y be a binomial random variable based on n trials and success probability θ. Then

$$\mu = E(Y) = n\theta, \quad \sigma^2 = V(Y) = n\theta(1-\theta).$$

The probability that the first success is to occur on the y-th trial:

Definition (Geometric distribution)

A random variable Y is said to have a geometric probability distribution with parameter θ if and only if

$$p(y) = (1-\theta)^{y-1}\theta, y = 1, 2, 3, ..., 0 \leq \theta \leq 1.$$

Theorem

If Y is a random variable with a geometric distribution,

$$\mu = E(Y) = \frac{1}{\theta}, \quad \sigma^2 = V(Y) = \frac{1-\theta}{\theta^2}.$$

Definition (Hypergeometric distribution)

A random variable Y is said to have a hypergeometric probability distribution if and only if $p(y) = \frac{C_y^r C_{n-y}^{N-r}}{C_n^N}$, where y is an integer 0, 1, 2, . . . , n, subject to the restrictions $y \le r$ and $n - y \le N - r$.

Theorem

If Y is a random variable with a hypergeometric distribution with parameter n, N and r, then we have

$$\mu = E(Y) = \frac{nr}{N}, \qquad \sigma^2 = V(Y) = n\frac{r}{N}\frac{N-r}{N}\frac{N-n}{N-1}.$$

For sampling without replacement, the number of successes in n trials is a random variable having a hypergeometric distribution with the parameters n, N and r.

Definition (Poisson distribution)

A random variable Y is said to have a Poisson distribution with *parameter* $\lambda > 0$ if

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \qquad y = 0, 1, 2, \ldots.$$

Theorem

If Y is a random variable possessing a Poisson distribution with parameter λ, then

$$\mu = E(Y) = \lambda, \quad \sigma^2 = V(Y) = \lambda.$$

Definition (Moment)

The k-th moment of a random variable Y taken about the origin is defined to be $E(Y^k)$ and is denoted by μ_k.

Definition (Moment generating function)

The moment-generating function m(t) for a random variable Y is defined to be $m(t) = E(e^{tY})$.

Theorem

If m(t) exists, then for any positive integer k, $\frac{d^k m(t)}{d^k t}|_{t=0} = \mu_k$. *In other words, if you find the k-th derivative of m(t) with respect to t and then set t = 0, the result will be* μ_k.

Assignment:

Find the moment-generating function $m(t)$ for a Poisson distributed random variable with mean λ.

Definition (Distribution function)

Let Y denote any random variable. The distribution function of Y , denoted by F(y), is such that F(y) $= P(Y \leq y)$ for $-\infty < y < \infty$.

Theorem (Properties of a Distribution Function)

If F(y) is a distribution function, then

- $F(-\infty) = \lim_{y\to-\infty} F(y) = 0$.
- $F(\infty) = \lim_{y\to\infty} F(y) = 1$.
- $F(y)$ *is a nondecreasing function of* y*. [If* y_1 *and* y_2 *are any values such that* $y_1 < y_2$*, then* $F(y_1) \leq F(y_2)$*.]*

Definition (Continuous distribution function)

A random variable Y with distribution function F(y) is said to be continuous if $F(y)$ is continuous, for $-\infty < y < \infty$.

Definition (Density function)

Let F(y) be the distribution function for a continuous random variable Y . Then f (y), given by

$$f(y) = \frac{dF(y)}{dy}$$

wherever the derivative exists, is called the probability density function for the random variable Y.

Theorem (Properties of a Density Function)

If f (y) is a density function for a continuous random variable, then

- $f(y) \geq 0$ *for all y.*
- $\int_{-\infty}^{\infty} f(y)dy = 1.$
- $F(y) = \int_{-\infty}^{y} f(x)dx.$

Theorem

If the random variable Y has density function $f(y)$ and $a < b$, then the probability that Y falls in the interval $[a, b]$ is

$$P(a \leq Y \leq b) = \int_a^b f(y)dy.$$

Definition (Expected value (or Expectation))

The expected value of a continuous random variable Y is

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy,$$

provided that the integral exists.

Theorem

Let g(Y) be a function of Y ; then the expected value of g(Y) is given by

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy,$$

provided that the integral exists.

Theorem (Linear property)

Let c be a constant and let g(Y), $g_1(Y), g_2(Y), ..., g_k(Y)$ *be functions of a continuous random variable* Y*. Then the following results hold:*

- $E[cg(Y)] = cE[g(Y)]$.
- $E[g_1(Y) + g_2(Y) + ... + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + ... + E[g_k(Y)]$.

Definition (Uniform distribution)

If $\theta_1 < \theta_2$, a random variable Y is said to have a continuous uniform probability distribution on the interval (θ_1, θ_2) if and only if the density function of Y is

$$f(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \le y \le \theta_2, \\ 0, & elsewhere. \end{cases}$$

Theorem

If $\theta_1 < \theta_2$ and Y is a random variable uniformly distributed on the interval (θ_1, θ_2), then

$$\mu = E(Y) = \frac{\theta_1 + \theta_2}{2}$$

and

$$\sigma^2 = V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}.$$

Definition (Normal distribution)

A random variable Y is said to have a normal probability distribution if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of Y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty.$$

- We denote this by $Y \sim N(\mu, \sigma^2)$.
- When $\mu = 0$ and $\sigma = 1$, it is called standard normal distribution.

Theorem

If Y is a normally distributed random variable with parameters μ and σ, then $E(Y) = \mu$ and $V(Y) = \sigma^2$.

Definition (Gamma distribution)

A random variable Y is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if and only if the density function of Y is

$$f(y) = \begin{cases} \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, & y \geq 0, \\ 0, & elsewhere, \end{cases}$$

where $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$.

Theorem

If Y has a gamma distribution with parameters α and β, then $\mu = E(Y) = \alpha\beta$ and $\sigma^2 = V(Y) = \alpha\beta^2$.

Definition (Chi-square distribution)

Let ν be a positive integer. A random variable Y is said to have a chi-square distribution with ν degrees of freedom if and only if Y is a gamma-distributed random variable with parameters $\alpha = \nu/2$ and $\beta = 2$. That is

$$f(y) = \begin{cases} \frac{y^{\nu/2-1}e^{-y/2}}{\beta^{\nu/2}\Gamma(\nu/2)}, & y \geq 0, \\ 0, & elsewhere. \end{cases}$$

Theorem

If Y is a chi-square random variable with degrees of freedom, then

$$\mu = E(Y) = \nu, \quad \sigma^2 = V(Y) = 2\nu.$$

If Y is a gamma-distributed random variable with parameters $\alpha = 1$:

Definition (Exponential distribution)

A random variable Y is said to have an exponential distribution with parameter $\beta > 0$ if and only if the density function of Y is $f(y) = \frac{1}{\beta} e^{-y/\beta}$ for $y \geq 0$ and $f(y) = 0$ elsewhere.

Theorem

If Y is an exponential random variable with parameter β, then

$$\mu = E(Y) = \beta, \quad \sigma^2 = V(Y) = \beta^2.$$

Example (Memoryless property)

Suppose that Y has an exponential probability density function. Show that, if $a > 0$ and $b > 0$,

$$P(Y > a + b | Y > a) = P(Y > b).$$

Proof.

Assignment! □

Definition (Moment)

If Y is a continuous random variable, then the kth moment about the origin is given by

$$\mu_k = E(Y^k), \qquad k = 1, 2, \ldots.$$

Definition (moment-generating function)

If Y is a continuous random variable, then the moment-generating function of Y is given by

$$m(t) = E(e^{tY}).$$

Theorem (Chebyshev's Theorem)

Let Y be a random variable with mean μ and finite variance σ^2. Then, for any constant $k > 0$,

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \quad \text{or} \quad P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Definition (Joint probability function)

Let Y_1 and Y_2 be discrete random variables. The joint (or bivariate) probability function for Y_1 and Y_2 is given by

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2), \quad -\infty < y_1 < \infty, -\infty < y_2 < \infty.$$

Theorem

If Y_1 and Y_2 are discrete random variables with joint probability function $p(y_1, y_2)$, then

- $p(y_1, y_2) \geq 0$ *for all* y_1, y_2;
- $\sum_{y_1, y_2} p(y_1, y_2) = 1$, *where the sum is over all values* (y_1, y_2) *that are assigned nonzero probabilities.*

Definition (Joint distribution function)

For any random variables Y_1 and Y_2, the joint (bivariate) distribution function $F(y_1, y_2)$ is

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2), \quad -\infty < y_1, y_2 < \infty.$$

Definition (Joint density function)

Let Y_1 and Y_2 be continuous random variables with joint distribution function $F(y_1, y_2)$. If there exists a nonnegative function $f(y_1, y_2)$, such that

$$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2) dt_2 dt_1,$$

for all $-\infty < y_1 < \infty, -\infty < y_2 < \infty$, then Y_1 and Y_2 are said to be jointly continuous random variables. The function $f(y_1, y_2)$ is called the joint probability density function.

Theorem

If Y_1 and Y_2 are random variables with joint distribution function $F(y_1, y_2)$, then

- $F(-\infty, -\infty) = F(-\infty, y_2) = F(y_1, -\infty) = 0.$
- $F(\infty, \infty) = 1.$
- *If $y_1' \geq y_1$ and $y_2' \geq y_2$, then*

$$F(y_1', y_2') - F(y_1, y_2') - F(y_1', y_2) + F(y_1, y_2) \geq 0.$$

Theorem

If Y_1 and Y_2 are jointly continuous random variables with a joint density function given by $f(y_1, y_2)$, then

- $f(y_1, y_2) \geq 0$ *for all* y_1, y_2.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1.$

Definition (Marginal probability distribution, Conditional probability distribution)

If Y_1 and Y_2 are jointly discrete random variables with joint probability function $p(y_1, y_2)$ and marginal probability functions $p_1(y_1)$ and $p_2(y_2)$, respectively, which are defined by

$$p_1(y_1) = P(Y_1 = y_1) = \sum_{all\ \ y_2} p(y_1, y_2),$$

$$p_2(y_2) = P(Y_2 = y_2) = \sum_{all\ \ y_1} p(y_1, y_2),$$

then the conditional discrete probability function of Y_1 given Y_2 is

$$p(y_1|y_2) = P(Y_1 = y_1|Y_2 = y_2) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p(y_1, y_2)}{p_2(y_2)},$$

provided that $p_2(y_2) > 0$.

Example (EXAMPLE 5.5)

From a group of three Republicans, two Democrats, and one independent, a committee of two people is to be randomly selected. Let Y_1 denote the number of Republicans and Y_2 denote the number of Democrats on the committee.

(a) Find the joint probability function of Y_1 and Y_2;

(b) Find the marginal probability function of Y_1;

(c) Find the conditional distribution of Y_1 given that $Y_2 = 1$, that is, given that one of the two people on the committee is a Democrat, find the conditional distribution for the number of Republicans selected for the committee.

Solution

Assignment!

Definition (Joint density function)

Let Y_1 and Y_2 be jointly continuous random variables with joint density $f(y_1, y_2)$ and marginal densities $f_1(y_1)$ and $f_2(y_2)$, respectively. For any y_2 such that $f_2(y_2) > 0$, the conditional density of Y_1 given $Y_2 = y_2$ is given by

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)}$$

and, for any y_1 such that $f_1(y_1) > 0$, the conditional density of Y_2 given $Y_1 = y_1$ is given by

$$f(y_2|y_1) = \frac{f(y_1, y_2)}{f_1(y_1)}.$$

Theorem

(1) If Y_1 and Y_2 are discrete random variables with joint probability function $p(y_1, y_2)$ and marginal probability functions $p_1(y_1)$ and $p_2(y_2)$, respectively, then Y_1 and Y_2 are independent if and only if

$$p(y_1, y_2) = p_1(y_1)p_2(y_2)$$

for all pairs of real numbers (y_1, y_2).

(2) If Y_1 and Y_2 are continuous random variables with joint density function $f(y_1, y_2)$ and marginal density functions $f_1(y_1)$ and $f_2(y_2)$, respectively, then Y_1 and Y_2 are independent if and only if

$$f(y_1, y_2) = f_1(y_1)f_2(y_2)$$

for all pairs of real numbers (y_1, y_2).

Example (EXAMPLE 5.10)

Refer to Example 5.5. Is the number of Republicans in the sample independent of the number of Democrats? (Is Y_1 independent of Y_2?)

Solution

Assignment!

Definition

Let $g(Y_1, Y_2, ..., Y_k)$ be a function of the discrete random variables, $Y_1, Y_2, ..., Y_k$, which have probability function $p(y_1, y_2, ..., y_k)$. Then the expected value of $g(Y_1, Y_2, ..., Y_k)$ is

$$E[g(Y_1, Y_2, ..., Y_k)] = \sum_{y_1,...,y_k} g(y_1, ..., y_k)p(y_1, ..., y_k).$$

If $Y_1, Y_2, ..., Y_k$ are continuous random variables with joint density function $f(y_1, y_2, ..., y_k)$, then

$$E[g(Y_1, Y_2, ..., Y_k)] = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} g(y_1, ..., y_k)p(y_1, ..., y_k)dy_1 ... dy_k.$$

Theorem

Let $g(Y_1, Y_2)$ be a function of the random variables Y_1 and Y_2 and let c be a constant. Then

$$E[cg(Y_1, Y_2)] = cE[g(Y_1, Y_2)].$$

Theorem

Let Y_1 and Y_2 be random variables and $g_1(Y_1, Y_2), ..., g_k(Y_1, Y_2)$ be functions of Y_1 and Y_2. Then

$$E[g_1(Y_1, Y_2)+...+g_k(Y_1, Y_2)] = E[g_1(Y_1, Y_2)]+...+E[g_k(Y_1, Y_2)].$$

Theorem

Let Y_1 and Y_2 be independent random variables and $g(Y_1)$ and $h(Y_2)$ be functions of only Y_1 and Y_2, respectively. Then

$$E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)],$$

provided that the expectations exist.

Definition (Covariance, Correlation coefficient)

If Y_1 and Y_2 are random variables with means μ_1 and μ_2, respectively, the covariance of Y_1 and Y_2 is

$$Cov(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)].$$

The correlation coefficient, ρ, a quantity related to the covariance and defined as

$$\rho = \frac{Cov(Y_1, Y_2)}{\sigma_1 \sigma_2},$$

where $\sigma_1^2 = V(Y_1)$ and $\sigma_2^2 = V(Y_2)$.

Theorem

If Y_1 and Y_2 are independent random variables, then

$$Cov(Y_1, Y_2) = 0.$$

Thus, independent random variables must be uncorrelated.

Theorem

Let $Y_1, Y_2, ..., Y_n$ and $X_1, X_2, ..., X_m$ be random variables with $E(Y_i) = \mu_i$ and $E(X_j) = \xi_j$. Define $U_1 = \sum_{i=1}^{n} a_i Y_i$ and $U_2 = \sum_{j=1}^{m} b_j X_j$ for constants $a_1, a_2, ..., a_n$ and $b_1, b_2, ..., b_m$. Then the following hold:

- $E(U_1) = \sum_{i=1}^{n} a_i \mu_i$.
- $V(U_1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j Cov(Y_i, Y_j)$.
- $Cov(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(Y_i, X_j)$.

Proof.

Assignment! □