STAT22000 Autumn 2013 Lecture 12&13

Yibi Huang

October 28, 2013

Random Variables

A **random variable** is a variable whose value depends on the outcome of a random phenomenon.

Formal Mathematical Definition of a Random Variable

A random variable is a function defined on the sample space $S = \{\text{all possible outcomes}\}$. The function assigns a value to each possible outcome.

Example 1. Let X be the number of heads in 3 tosses. Then $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ and

$$X(HHH) = 3, \quad X(HHT) = 2, \quad X(HTH) = 2, \quad X(HTT) = 1,$$
$$X(THH) = 2, \quad X(THT) = 1, \quad X(TTH) = 1, \quad X(TTT) = 0$$

Example 2. Let Y be the number of toss required to get a head. Then $S = \{H, TH, TTH, TTTH, TTTTH, \ldots\}$ and

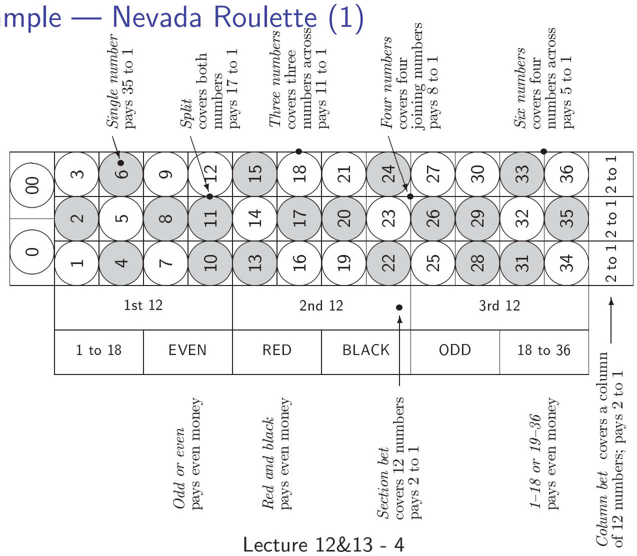
$$Y(H) = 1,\ Y(TH) = 2,\ Y(TTH) = 3,\ Y(TTTH) = 4, \ldots$$

Discrete Random Variables

- A **discrete** random variable takes on finitely many possible values
- A **distribution** of a discrete random variable is a list of its possible values and the probabilities that it takes on those values.
- e.g., X = number of heads in 3 tosses

Value of X	0	1	2	3
Outcomes	TTT	HTT THT TTH	HHT HTH THH	HHH
Probability	1/8	3/8	3/8	1/8

Example — Nevada Roulette (1)



Example — Nevada Roulette (2)

A gambler is going to make two bets:

- a *single* bet at number 1 for one dollar, and
- a *split* bet at number 2 and 3, for another dollar

- If 1 comes up, the gambler wins the single bet. He gets the dollar back, together with winnings of \$35. Otherwise he loses his dollar on the single bet.
- If either 2 or 3 comes up, the gambler wins the split bet. He gets the dollar back, together with winnings of \$17. If neither number comes up, he loses his dollar on the split bet.

Example — Nevada Roulette (3)

Roulette outcomes	X	Y	T
0	−1	−1	−2
00	−1	−1	−2
1	35	−1	34
2	−1	17	16
3	−1	17	16
4	−1	−1	−2
5	−1	−1	−2
6	−1	−1	−2
7	−1	−1	−2
8	−1	−1	−2
9	−1	−1	−2
10	−1	−1	−2
⋮	⋮	⋮	⋮
35	−1	−1	−2
36	−1	−1	−2

Let X be the earning of the gambler on the single bet, Y be his earning on the split bet, and $T = X + Y$ be the total earning on both bets.

Distribution of X:

value of X	35	−1
probability	$\frac{1}{38}$	$\frac{37}{38}$

Distribution of Y:

value of Y	17	−1
probability	$\frac{2}{38}$	$\frac{36}{38}$

Distribution of $T = X + Y$:

value of T	34	16	−2
probability	$\frac{1}{38}$	$\frac{2}{38}$	$\frac{35}{38}$

Example — Sampling From a Mini-Population (1)

An investigator wants to study a population with 4 individuals only. He wants to estimate two *parameters*:

p = fraction of population who vote for candidate A, and
μ = average age of the population

Unknown to the investigator, the 4 individuals in the population are

individual	age	vote for
Adam	20	candidate A
Betty	30	candidate A
Clare	40	candidate B
David	50	candidate B

Here $p = 2/4 = 0.5$, and

$\mu = \frac{20+30+40+50}{4} = 35$

If the investigator takes a simple random sample (SRS) of size 2, and estimates p and μ by

$\widehat{p}$ = fraction of the sample voting for candidate A, and

$\widehat{\mu}$ = average age of the sample,

what is the distribution of $\widehat{p}$ and $\widehat{\mu}$?

Example — Sampling From a Mini-Population (2)

Let's list all possible SRS and the corresponding estimates.

Population

individual	age	vote
Adam	20	A
Betty	30	A
Clare	40	B
David	50	B

$\Rightarrow$

Sample	Votes	Ages	$\widehat{p}$	$\widehat{\mu}$
Adam & Betty	AA	20,30	1	25
Adam & Clare	AB	20,40	0.5	30
Adam & David	AB	20,50	0.5	35
Betty & Clare	AB	30,40	0.5	35
Betty & David	AB	30,50	0.5	40
Clare & David	BB	40,50	0	45

The distribution of $\widehat{p}$ and $\widehat{\mu}$ are respectively:

value of $\widehat{p}$	0	0.5	1
probability	1/6	4/6	1/6

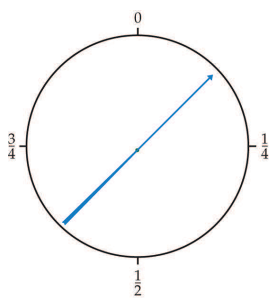
, and

value of $\widehat{\mu}$	25	30	35	40	45
probability	1/6	1/6	2/6	1/6	1/6

In sampling, $\widehat{p}$ and $\widehat{\mu}$ are called **statistics**, and their distributions are called the **sampling distributions**.

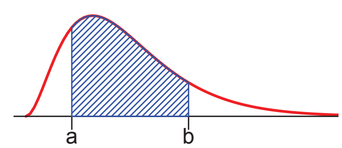
Example of a Continuous Random Variable

- A spinner turns freely on its axis and slowly comes to a stop.
- Define a random variable X as the location of the pointer when the spinner stops. It can be anywhere on a circle that is marked from 0 to 1.
- Sample space $S = \{$ all numbers x such that $0 \leq x < 1\}$
- $P(0.3 < X < 0.7) = ?$
- $P(X < 0.5 \text{ or } X > 0.8) = ?$
- $P(X = 0.75) = ?$



Continuous Random Variables

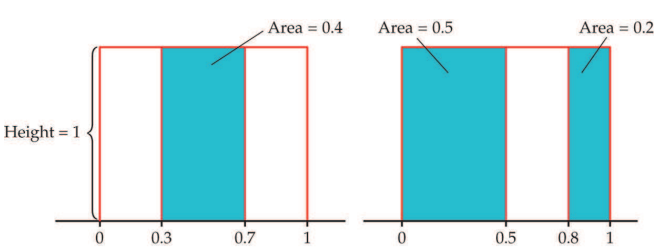
- A continuous random variable takes all values in an interval of numbers
  - Note: the interval does not have to be bounded
- The probability distribution of a continuous random variable is described by a **density curve**.
- A density curve stays above 0 and the total area under it is 1.
- If Y is a continuous random variable, $P(a < Y < b)$ is the area under the density curve of Y above the interval between a and b



- Note: all continuous probability distributions assign zero probability to every individual outcome: $P(Y = y) = 0$

Spinner Example Revisit

For the spinner example, the density curve for X is constant at 1 on the interval $[0, 1]$, and 0 elsewhere.



$P(0.3 < X < 0.7) = 0.4$ $\quad$ $P(X < 0.5 \text{ or } X > 0.8)$

Independent Random Variables

- Idea: knowing information about the value of X tells us nothing about the value of Y.
- Two discrete random variables X and Y are independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all numbers x and y. i.e.

$$P(X = x \text{ and } Y = y) = P(X = x)\,P(Y = y)$$

- Two continuous random variables X and Y are independent means that the events $\{a < X < b\}$ and $\{c < Y < d\}$ are independent for all numbers a, b, c, and d. i.e.

$$P(a < X < b \text{ and } c < Y < d)$$
$$= P(a < X < b)\,P(c < Y < d)$$

  - i.e., the multiplication rule

For the Roulette example,

$$P(X = 35,\ Y = 17) = 0 \neq P(X = 35)P(Y = 17) = \frac{1}{38} \times \frac{2}{38},$$

so X and Y are not independent.

That is not surprising.
If the gambler wins his single bet at 1, he must lose his split bet at 2 and 3. The two bets are dependent.

For the sampling from mini-population example, are $\widehat{p}$ and $\widehat{\mu}$ independent?

$$P(\widehat{p} = 0) = \frac{1}{6}$$
$$P(\widehat{\mu} = 25) = \frac{1}{6}$$
$$P(\widehat{p} = 0 \text{ and } \widehat{\mu} = 25) = ?$$

Mean of a Random Variable

For a discrete random variable X with probability distribution

value of X	x_1	x_2	$\cdots$	x_n
probability	p_1	p_2	$\cdots$	p_n

the **mean** of X (or the **expected value** of X) is found by multiplying each possible value of X by its probability, and then adding the products.

$$\mu_X = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_i x_i p_i$$

Notation:

$$\begin{aligned}\text{mean of } X &= \text{expected value of } X \\ &= \mu_X = \mu(X)\end{aligned}$$

Nevada Roulette Example Revisit

Distribution of X:

value of X	35	-1
probability	$\frac{1}{38}$	$\frac{37}{38}$

$\Rightarrow\ \mu_X = 35 \times \frac{1}{38} + (-1) \times \frac{37}{38} = -\frac{2}{38}$

Distribution of Y:

value of Y	17	-1
probability	$\frac{2}{38}$	$\frac{36}{38}$

$\Rightarrow\ \mu_Y = 17 \times \frac{2}{38} + (-1) \times \frac{36}{38} = -\frac{2}{38}$

Distribution of $T = X + Y$:

value of T	34	16	-2
probability	$\frac{1}{38}$	$\frac{2}{38}$	$\frac{35}{38}$

$\Rightarrow\ \mu_T = 34 \cdot \frac{1}{38} + 16 \cdot \frac{2}{38} + (-2)\frac{35}{38} = -\frac{4}{38}$

Observe that $\mu_T = \mu_X + \mu_Y$.

Why is the Mean Defined This Way?

This definition makes the following law hold:

Law of Large Numbers (LLN)

As we do many independent repetitions of the experiment, drawing more and more observations from the same distribution, the sample mean will approach the mean of the distribution more and more closely.

Law of Large Numbers for the Spinner Example

Imagine the gambler playing roulette n times, using the same betting strategy (a single at 1 and a split at 2, 3), with n large. Let $T_1, T_2, \ldots, T_n$ be the respective winnings of the 1st, 2nd,..., nth play.

The total winnings $T_1 + T_2 + \cdots + T_n$ will be

$$34 \times (\#\text{ of 34's}) + 16 \times (\#\text{ of 16's}) + (-2) \times (\#\text{ of (-2)'s}).$$

Then the average winnings per play $\overline{T}_n = (T_1 + T_2 + \cdots + T_n)/n$ is

$$34 \times \underbrace{\left(\frac{\#\text{ of 34's}}{n}\right)}_{\downarrow\ \frac{1}{38}} + 16 \times \underbrace{\left(\frac{\#\text{ of 16's}}{n}\right)}_{\downarrow\ \frac{2}{38}} + (-2) \times \underbrace{\left(\frac{\#\text{ of (-2)'s}}{n}\right)}_{\downarrow\ \frac{35}{38}}$$

I.e., for large n

$$\overline{T}_n \to 34 \times \frac{1}{38} + 16 \times \frac{2}{38} + (-2) \times \frac{35}{38} = -\frac{4}{38} = \mu_T.$$

Mini-Sampling Example Revisit

Distribution of $\widehat{p}$:

value of $\widehat{p}$	0	0.5	1
probability	1/6	4/6	1/6

The mean of $\widehat{p}$ is

$$0 \times \frac{1}{6} + 0.5 \times \frac{4}{6} + 1 \times \frac{1}{6} = 0.5 = p$$

The mean of $\widehat{p}$ is the same as the parameter p we want to estimate. We say $\widehat{p}$ is an **unbiased** estimator of p.

Distribution of $\widehat{\mu}$:

value of $\widehat{\mu}$	25	30	35	40	45
probability	1/6	1/6	2/6	1/6	1/6

The mean of $\widehat{\mu}$ is

$$25 \cdot \frac{1}{6} + 30 \cdot \frac{1}{6} + 35 \cdot \frac{2}{6} + 40 \cdot \frac{1}{6} + 45 \cdot \frac{1}{6} = 35 = \mu$$

$\widehat{\mu}$ is also an **unbiased** estimator of μ.

Mean for A Continuous Random Variables (1)

If X is a **continuous** random variable with density curve $f(x)$. The mean of X is defined as the integral

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx$$

For example, for the spinner example, the density of X is a constant 1 on [0,1] and 0 elsewhere

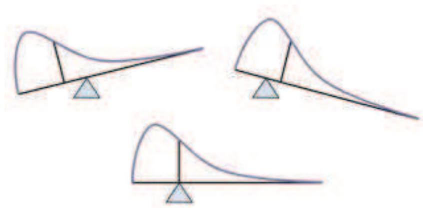
$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x > 1 \end{cases}$$

The mean of X is

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{1} x \cdot 1dx = \frac{1}{2}x^2\Big|_0^1 = \frac{1}{2}.$$

Mean for A Continuous Random Variables (2)

Recall in Lecture 3 we say the mean of a density curve is the **balance point** of the curve.



Here we define the mean of a density curve as

$$\int_{-\infty}^{\infty} xf(x)dx$$

These two definitions are equivalent.

Variances for Discrete Random Variables

For a discrete random variable X with probability distribution

value of X	x_1	x_2	$\cdots$	x_n
probability	p_1	p_2	$\cdots$	p_n

the **variance** σ_X^2 of X is found by multiplying each squared deviation of X by its probability and then adding all the products.

$$\begin{aligned}\sigma_X^2 &= (x_1-\mu_X)^2p_1 + (x_2-\mu_X)^2p_2 + \cdots + (x_n-\mu_X)^2p_n \\ &= \sum_i (x_i-\mu_X)^2p_i\end{aligned}$$

Notation:

$$\begin{aligned}\text{variance of } X &= \sigma_X^2 = \sigma^2(X) \\ \text{SD of } X &= \sqrt{\text{variance of } X} = \sigma_X = \sigma(X)\end{aligned}$$

Nevada Roulette Example Revisit

Distribution of X:

value of X	35	-1
probability	$\frac{1}{38}$	$\frac{37}{38}$

We have found that $\mu_X = -1/19$. So the variance of X is

$$\begin{aligned}\sigma_X^2 &= [35-(-\frac{1}{19})]^2 \times \frac{1}{38} + [-1-(-\frac{1}{19})]^2 \times \frac{37}{38} \\ &= \left(35\frac{1}{19}\right)^2 \times \frac{1}{38} + \left(\frac{18}{19}\right)^2 \times \frac{37}{38} \\ &= \frac{11988}{361} = \frac{18^2 \times 37}{19^2} \approx 33.21\end{aligned}$$

and the SD of X is

$$\sigma_X = \sqrt{\sigma_X^2} = \frac{18\sqrt{37}}{19} \approx 5.763.$$

An Alternative Formula for Variance

Observe that

$$\begin{aligned}\sigma_X^2 &= \sum\nolimits_i (x_i-\mu_X)^2p_i \\ &= \sum\nolimits_i (x_i^2 - 2\mu_X x_i + \mu_X^2)p_i \\ &= \sum\nolimits_i x_i^2p_i - 2\sum_i \mu_X x_i p_i + \sum_i \mu_X^2 p_i \\ &= \sum\nolimits_i x_i^2p_i - 2\mu_X \underbrace{\sum\nolimits_i x_i p_i}_{=\mu_X} + \mu_X^2 \underbrace{\sum\nolimits_i p_i}_{=1} \\ &= \sum\nolimits_i x_i^2p_i - 2\mu_X \cdot \mu_X + \mu_X^2 \\ &= \sum\nolimits_i x_i^2p_i - \mu_X^2\end{aligned}$$

$$\boxed{\sigma_X^2 = \sum\nolimits_i x_i^2p_i - \mu_X^2}$$

Nevada Roulette Example Revisit

Distribution of X:

value of X	35	-1
probability	$\frac{1}{38}$	$\frac{37}{38}$

$\mu_X = -\frac{1}{19}.$

Using the alternative formula, the variance of X is

$$\begin{aligned}\sigma_X^2 &= 35^2 \times \frac{1}{38} + (-1)^2 \times \frac{37}{38} - \mu_X^2 \\ &= \frac{35^2+37}{38} - \frac{1}{19^2} \\ &= \frac{11988}{361} \approx 33.21\end{aligned}$$

Nevada Roulette Example Revisit

Recall the distributions of Y and T are respectively:

value of Y	17	-1
probability	$\frac{2}{38}$	$\frac{36}{38}$

value of T	34	16	-2
probability	$\frac{1}{38}$	$\frac{2}{38}$	$\frac{35}{38}$

and their means are $\mu_Y = -\frac{1}{19}$, and $\mu_T = -\frac{2}{19}$.
Using the alternative formula, the variance of Y is

$$\begin{aligned}\sigma_Y^2 &= 17^2 \times \frac{2}{38} + (-1)^2 \times \frac{36}{38} - \mu_Y^2 \\ &= \frac{17^2 \times 2 + 36}{38} - \frac{1}{19^2} = \frac{5832}{361} \approx 16.155,\end{aligned}$$

and the variance of T is

$$\begin{aligned}\sigma_T^2 &= 34^2 \times \frac{1}{38} + 16^2 \times \frac{2}{38} + (-2)^2 \times \frac{35}{38} - \mu_T^2 \\ &= \frac{34^2 + 16^2 \times 2 + (-2)^2 \times 35}{38} - \frac{(-2)^2}{19^2} = \frac{17172}{361} \approx 47.568\end{aligned}$$

Variance of Continuous Random Variables

If X is a **continuous** random variable with density curve $f(x)$. The variance of X is defined as the integral

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx$$

in which μ_X is the mean of X.

There is also an alternative formula for the variance of continuous random variables.

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu_x^2$$

Variance for the Spinner Example

For the spinner example, recall the density of X is a constant 1 on [0,1] and 0 elsewhere

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x > 1 \end{cases}$$

and $\mu_X = 0.5$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-0.5)^2 f(x) dx = \int_0^1 (x-0.5)^2 \cdot 1 dx = \frac{1}{3}(x-0.5)^3\Big|_0^1 = \frac{1}{12}.$$

or alternatively,

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu_X^2 = \int_0^1 x^2 \cdot 1 dx - \mu_X^2 = \frac{1}{3}x^3\Big|_0^1 - (0.5)^2 = \frac{1}{12}.$$

Properties of Mean and Variance

Suppose X is a random variable and c is a fixed number. Then

- $\mu(X + c) = \mu(X) + c$, $\mu(cX) = c\mu(X)$
- $\sigma(X + c) = \sigma(X)$
- $\sigma(cX) = |c|\sigma(X)$, $\sigma^2(cX) = c^2\sigma^2(X)$

Suppose X and Y are random variables. Then

- $\mu(X + Y) = \mu(X) + \mu(Y)$ (always valid)
- $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$ when X and Y are independent

Question: What about $\sigma^2(X - Y)$?

For the Roulette example, as $T = X + Y$, we have

$$\mu_T = -\frac{2}{19} = \mu_X + \mu_Y = -\frac{1}{19} + (-\frac{2}{19})$$

but

$$\sigma_T^2 \approx 47.568 < \sigma_X^2 + \sigma_Y^2 \approx 33.21 + 16.155 \approx 49.365$$

since X and Y are NOT independent.

Exercise: Coin Tossing (1)

Toss a coin 3 times. Let

$X_1 = 1$ number of heads in the first two tosses
$X_2 = 1$ if getting a head in the third toss, and 0 if tails,
$S = 1$ number of heads in the three tosses

Observe that $S = X_1 + X_2$.

Show that the distributions of X_1, X_2, S are respectively

value of X_1	0	1	2
probability	1/4	1/2	1/4

value of X_2	0	1
probability	1/2	1/2

and

value of S	0	1	2	3
probability	1/8	3/8	3/8	1/8

Exercise: Coin Tossing (2)

$$\begin{aligned}\mu_{X_1} &= 0 \cdot (1/4) + 1 \cdot (1/2) + 2 \cdot (1/4) = 1 \\ \mu_{X_2} &= 0 \cdot (1/2) + 1 \cdot (1/2) = 1/2 \\ \mu_S &= 0 \cdot (1/8) + 1 \cdot (3/8) + 2 \cdot (3/8) + 3 \cdot (1/8) = 3/2\end{aligned}$$

Observe that $\mu_{X_1} + \mu_{X_2} = \mu_S$.

$$\begin{aligned}\sigma_{X_1}^2 &= 0^2 \cdot (1/4) + 1^2 \cdot (1/2) + 2^2 \cdot (1/4) - \mu_{X_1}^2 = 1/2 \\ \sigma_{X_2}^2 &= 0^2 \cdot (1/2) + 1^2 \cdot (1/2) - \mu_{X_2}^2 = 1/4 \\ \sigma_S^2 &= 0^2 \cdot (1/8) + 1^2 \cdot (3/8) + 2^2 \cdot (3/8) + 3^2 \cdot (1/8) - \mu_S^2 = 3/4\end{aligned}$$

Observe that $\sigma_{X_1}^2 + \sigma_{X_2}^2 = \sigma_S^2$.
This is true because the outcome of the third toss is independent of the outcome of the first two tosses, i.e., X_1 and X_2 are independent.