Math 425
Introduction to Probability
Lecture 14

Kenneth Harris
kaharri@umich.edu

Department of Mathematics
University of Michigan

February 9, 2009

Statistics and Random Variables

☞ When a large collection of numbers is assembled, as in a census, we are usually not as interested in the individual values, but in a summary of the data.

We use descriptive quantities such as mean, median and deviation to help summarize data.

☞ The same is true for random variables. We will be mainly concerned with two such descriptive quantities.

- The expected value (or mean) of the random variable provides a weighted average of the possible values of the random variable. It gives you the central tendancy of values.
- The variance of the random variable provides a measure of the spread of the possible values of the random variable around its expected value.

Finite valued Random Variables

☞ We will start out with finite discrete random variables.

☞ Let X be a random variable with probability mass function p_X.
X is finite if the set of possible values of X

$$D_X = \{r \in \mathbb{R} \,|\, p_X(r) > 0\} \quad \text{is finite .}$$

Expectation defined

Definition

Let X be a finite random variable with probability mass function p_X and possible values D_X.

The expected value of X is defined by

$$E[X] = \sum_{r \in D_X} r \cdot p_X(r).$$

This value is also called the mean or average of the random variable.

Note. Ross defines expectation in Chapter 4 over arbitrary discrete random variables. We will discuss infinite discrete random variables in a couple lectures.

Example: Fair coin tosses

Example. Let X be the random variable which counts the number of heads in a single toss of a fair coin. It has pmf:

$$p(0) = \frac{1}{2} = p(1)$$

and expectation

$$E[X] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}.$$

Note. The mean need not be an actual value.

☺ How many times have you heard the less statistically-minded have a good time with the “average family size" with fractional children. Of course, real children may be fractious, but never fractional.

Example: Fair coin tosses

Example. Let X be the random variable which counts the number of heads in two fair coin tosses. Then X has pmf

$$p(0) = \frac{1}{4} \quad p(1) = \frac{1}{2} \quad p(2) = \frac{1}{4},$$

and expectation

$$E[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Example: die

Example. Let X be the random variable which counts the face value of a single thrown die. Then X has pmf

$$p(n) = \frac{1}{6} \qquad 1 \leq n \leq 6,$$

and expectation

$$\begin{aligned} E[X] &= \sum_{n=1}^{6} n \cdot \frac{1}{6} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5. \end{aligned}$$

Example: pairs of dice

Example. Let X be the random variable which counts the face value of a pair of thrown dice. Then X has pmf

n	$p(n)$	n	$p(n)$
2	$\frac{1}{36}$	7	$\frac{6}{36}$
3	$\frac{2}{36}$	8	$\frac{5}{36}$
4	$\frac{3}{36}$	9	$\frac{4}{36}$
5	$\frac{4}{36}$	10	$\frac{3}{36}$
6	$\frac{5}{36}$	11	$\frac{2}{36}$
		12	$\frac{1}{36}$

and expectation

$$E[X] = \sum_{n=2}^{12} n \cdot p(n) = 7.$$

Example: Indicator variables

Example. Let S be a sample space and $A \subseteq S$ an event.
Define a random variable, I_A (the indicator variable for E) for each $s \in S$:

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A. \end{cases}$$

The pmf for I_E is

$$p(1) = \mathbf{P}(A) \qquad p(0) = 1 - \mathbf{P}(A)$$

and the expectation is

$$E[I_A] = 1 \cdot \mathbf{P}(A) + 0 \cdot (1 - \mathbf{P}(A)) = \mathbf{P}(A).$$

Mean and a fallacy

☞ The mean of a random variable provides a description of the central tendancy of trials over the long run. (We will make this precise with the Strong Law of Large Numbers in Chapter 8.)

☠ Warning. Many people would agree with the following statement

> If the average lifespan is 75 years, then it is an evens chance that any newborn will live for more than 75 years.

But this is NOT TRUE. The mean tells you nothing about what is likely to happen in any single case.

Examples

☞ It is not generally true for a random variable X that

$$\sum_{r > E[X]} p_X(r) = \frac{1}{2}.$$

Example. The expected value of a bet in chuck-a-luck was 0, although 58% of all throws of the dice were losses. See Lecture 13.
Example. Let X be a random variable with the following pmf:

$$p(-100) = 0.99 \qquad p(10000) = 0.01,$$

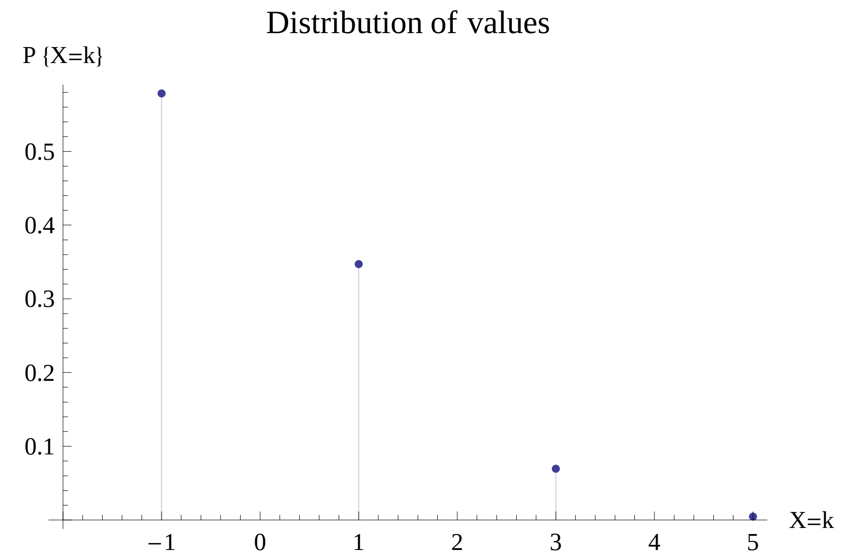
and expectation

$$E[X] = (-100) \cdot 0.99 + 10000 \cdot 0.01 = 1.$$

The mean is a weighted average that can be skewed by high-valued, low-probability outcomes.

Distribution of values for Chuck-a-Luck

☞ Distribution of values for chuck-a-luck.



Median and mode

☞ Outcomes are equally likely to occur on either side of the median, and the most likely outcomes are modes.

Definition

Let X be a random variable with pmf $p(x)$ and possible values D_X.
If m is any number such that

$$\sum_{r \le m} p(r) \ge \frac{1}{2} \quad \text{and} \quad \sum_{r \ge m} p(r) \ge \frac{1}{2}$$

then m is a median of X.

If $\lambda \in D_X$ is such that

$$p(\lambda) \ge p(r) \qquad \text{for any } r \in D_X$$

then λ is said to be a mode of X.

Median and mode

Example. Let X be a random variable with pmf

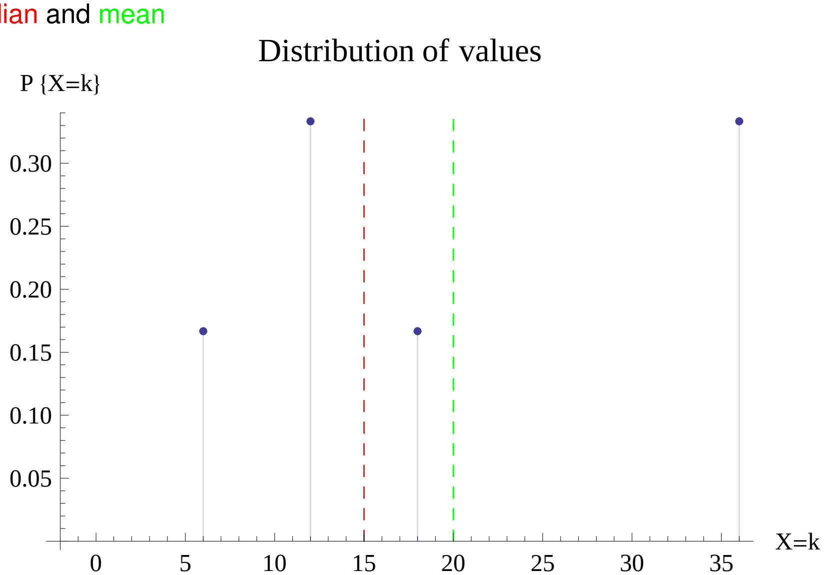
$$p(6) = \frac{1}{6} \quad p(12) = \frac{2}{6} \quad p(18) = \frac{1}{6} \quad p(36) = \frac{2}{6}$$

The mean is 20, a median is 15 and there are two modes, 12 and 36.

☞ The mean and mode will be close when there is not too much spread in the data. We will want another statistic which provides a measure of spread from the mean, the variance (next lecture).

Distribution of values

median and mean



Example: Bendford distribution

Example. Last lecture we saw that in the leading digit $\{1, 2, \ldots, 9\}$ in many data sets that occur in “real life" (such as census data) are not uniformly distributed (meaning that each digit is equally likely to occur), but are distributed by

$$p(n) = \log_{10}(1 + \frac{1}{n}) \qquad 1 \le n \le 9.$$

(This is caled the Benford distribution.)

The expected value of the leading first digit is

$$\sum_{n=1}^{9} n \cdot p(n) \approx 3.44.$$

What does this mean?

Sample mean and relative frequency

☞ Suppose we have sampled a large number N from a population of similar items, such as a large data set. We measure some numerical quantity, such as the leading digit from each data item. In this way we obtain a collection of observations:

$$O = (x_1, x_2, \ldots, x_N).$$

☞ The sample mean of the observations is

$$\overline{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i.$$

Sample mean and relative frequency

☞ Let $\#(x)$ be the number of times the value x occurred in the data set

$$O = (x_1, x_2, \ldots, x_N).$$

The relative frequency of x in the observations is

$$P(x) = \frac{\#(x)}{N}$$

We can restate the sample mean of the observations as

$$\overline{x} = \sum_{x \in O} x \cdot P(x).$$

Sample mean and relative frequency

☞ A statistician proposes a mathematical model to explain the data. In this case, the statistician proposes that that the measured outcomes, given by a random variable X, has the following probability mass function:

$$p(n) = \log_{10}(1 + \frac{1}{n}) \qquad 1 \le n \le 9.$$

and expected value

$$E[X] = 3.44$$

Sample mean and relative frequency

☞ What is the relation between the statisticians expected value $E[X]$ and the sample mean of the data $\overline{x}$?

☞ If the statistician's model is accurate then for a large sample size N, the relative frequency $P(x)$ is close to the probability $p(x)$ for each possible value x. So,

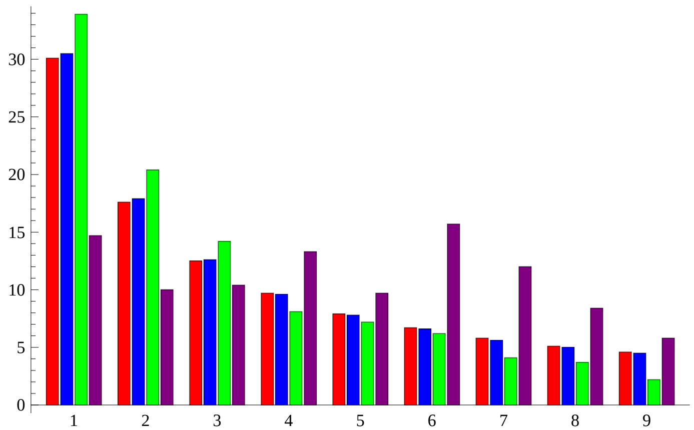
$$\overline{x} = \sum_{x} x \cdot P(x) \approx \sum_{x} x \cdot p(x) = E[X].$$

That is, the sample mean $\overline{x}$ will closely approximate the expected value $E[X]$.

Sample mean and relative frequency

☞ Benford's distribution arises from data sets which are "naturally occurring".

Benford, Tax Returns, Population, Random



Example: 16 Questions

Example

A number is chosen at random between 1 and 16. You are to try to guess the number by asking "yes-no" questions. Here are two strategies.

(a) For each i from 1 to 16 you ask: "Is it i?".

(b) Pick the midpoint M of the known range and ask: "Is it bigger than M?"

You first question: " Is it bigger than 8.5?"

- If Yes, ask: "Is it bigger than $\frac{9+16}{2} = 12.5$?";
- If No, ask: "Is it bigger than $\frac{1+8}{2} = 4.5$?"

(This cuts the range of possible values in half each time.)

What is the expected number of questions?

Example: 16 Questions

(b). Each question cuts the range of values in half; so, by asking $4 = \log_2 16$ questions you reduce the possible numbers to 1.

5 questions suffice each time.

☞ Suppose you guess when there are two possibilities.

On the fourth question you guess on the two remaining values. So, half the time you need 4 questions and half the time 5 questions.

The expected number of questions is now 4.5 questions.

Example: 16 Questions

(a). Let X be a random variable which counts the number of questions asked when the correct number was guessed.

Intuitively (I hope), you should expect

$$\mathbf{P}\{X = i\} = \frac{1}{16} \qquad 1 \le i \le 16.$$

☞ Here is the formal nonsense:

$$\mathbf{P}\{X = i\} = \frac{15 \cdot 14 \cdot (15 - i + 1)}{16 \cdot 15 \cdot (16 - i + 1)} \cdot \frac{1}{16 - i} = \frac{1}{16}$$

Example: 16 Questions

☞ The expected number of questions in strategy **(a)** is

$$E[X] = \frac{1}{16}\left(\sum_{i=1}^{16} i\right) = 8.5.$$

☞ Strategy **(b)** requires only 5 questions, so is the better choice – over the long run.

☞ In fact, we can never expect to do better than strategy **(b)** (which is known as the divide-and-conquer strategy).

Example: Playoff

Example

The Red Wings and Blackhawks are playing a best of three format (first team to win two games wins the series.) If the Red Wing's probability of winning any game is p (independently of the previous games), what is the expected number of games to be played?

Example: Playoff

☞ The playoff can go 2 or 3 games with either the Red Wings or Black Hawks winning. Use the following random variable

- $\{X = i\}$: the playoff goes i games (so, $i = 2, 3$ are the only values that have nonzero probability).

☞ To compute the probabilities we need to compute the probability that each team wins the playoff in that number of games.

$$\begin{aligned} \mathbf{P}\{X=2\} &= p^2 + (1-p)^2 \quad \text{Wings win + Hawks win} \\ &= 1 + 2p^2 - 2p \\ \mathbf{P}\{X=3\} &= \binom{2}{1}\cdot p^2(1-p) + \binom{2}{1}\cdot p(1-p)^2 \quad \text{Wings win + Hawks win} \\ &= 2p - 2p^2 \end{aligned}$$

Note: in the case of $X = 3$ the winner of the third game is the series winner.

Example: Playoff

☞ So, the expected number of games is

$$\begin{aligned} E[X] &= 2\cdot\mathbf{P}\{X=2\} + 3\cdot\mathbf{P}\{X=3\} \\ &= 2(1+2p^2-2p) + 3(2p-2p^2) \\ &= 2 + 2p - 2p^2 = 2(1+p-p^2) \end{aligned}$$

Examples.

- If $p = \frac{1}{4}$, then $E[X] = \frac{19}{8} = 2.375$.
- If $p = \frac{1}{2}$, then $E[X] = \frac{5}{2} = 2.5$.
- The expectation $E[X] = 2(1 + p - p^2)$ has a local maximum at $\frac{1}{2}$, so the maximum number of expected games is 2.5.

Example: Betting on evens

Example

A die is thrown. If an odd number turns up, we win an amount equal to this number; if an even number turns up we lose an amount equal to this number.

☞ Let X be a random which counts our winnings.
The expected value of X is

$$E[X] = 1 \cdot \frac{1}{6} - 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} - 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} - 6 \cdot \frac{1}{6} = -0.5$$

☞ Over many bets, we would expect to lose, on average, 50 cents per bet.
If we place 10,000 bets this amounts to a loss of $5000.

Example: Betting on evens

☞ I placed this bet over 10,000 plays and computed the sample mean.
This was repeated 1000 times with the following results:

- mean = -0.5033
- maximum = -0.4194
- minimum = -0.6129
- standard deviation = 0.0364

☞ This agrees quite well with the prediction of the mathematical model.

Example: Fair games

☞ Consider a game (i.e. an experiment) with outcomes $\{x_1, x_2, \ldots, x_n\}$.
A bet is a random variable that on outcome x_i pays $X(x_i)$:

- If $X(x_i) > 0$, you make money on the bet.
- If $X(x_i) < 0$, you lose money on the bet.

☞ A bet X for which $E[X] = 0$ is said to be fair, and otherwise unfair.
An unfair bet is

- An unfair bet is favorable if $E[X] > 0$ (you expect to make money),
- An unfair bet is unfavorable if $E[X] < 0$ (you expect to lose money).

☺ What is “favorable" to the bettor is “unfavorable" to the house.

Example: Heads

Example. A dollar bet on heads of a fair coin toss. The random variable X representing your winnings is

$$E[X] = 1 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} = 0.$$

This is an fair game.

Example. A dollar bet on heads from a biased coin which favors tails at $\frac{3}{4}$. The random variable X representing your winnings is

$$E[X] = 1 \cdot (\frac{1}{4}) + (-1) \cdot (\frac{3}{4}) = -\frac{1}{2}$$

This is an unfair game that costs 50 cents on each dollar bet.

Example: Craps

Example. In the game of craps, the probability that the shooter wins is about 0.493 (from Lecture 12).

A bet X at even money has expected value

$$E[X] = 1 \cdot (0.493) + (-1) \cdot (0.507) \approx -0.014.$$

This is an unfair game, and you can expect to lose about 1.5 cents over each play.

Example: Pick-six Lotto

Example. In Pick-six lotto you try to match six numbers from $\{1, \ldots, 49\}$. We calculated the odds (Lecture 10) at $\frac{1}{13{,}983{,}816}$.

Suppose you pay \$1 for a ticket with a ten \$10,000,000 (no ties). The expectation of this bet X is

$$\begin{aligned} E[X] = &= 10^6 \cdot \left(\frac{1}{13,983,816}\right) + (-1) \cdot \left(1 - \frac{1}{13,983,816}\right) \\ &\approx -0.2849 \end{aligned}$$

This is an unfair game, and you can expect to lose about 29 cents on each dollar.

Example: Roulette

Example. A roulette wheel is divided into 38 slotted sectors.

- 18 are red, 18 are black and 2 are green,
- 36 are numbered 1 to 36, together with one marked 0 and one marked 00.

People often bet even money on color (red/black) or parity (even/odd).

Are these fair bets?

$$\mathbf{P}(red) = \mathbf{P}(black) = \mathbf{P}(odd) = \mathbf{P}(even) = \frac{18}{38} \approx 0.474.$$

The bet X of \$1 on red has expected value

$$E[X] = 1 \cdot (0.474) + (-1) \cdot (.526) = -0.052.$$

This bet is unfair and you can expect to lose 5 cents per bet over the long run.

Example: Roulette

☞ Another bet on roulette is on a particular number (1 to 36), with a Vegas payoff of \$35 on a \$1 bet. The expected value of this bet is

$$35 \cdot \frac{1}{38} + (-1) \cdot \frac{37}{38} = -\frac{2}{38} \approx -0.0526$$

An unfair bet for you.

☞ A fair payoff for this bet is \$37 on a \$1 bet:

$$E[X] = 37 \cdot \frac{1}{38} + (-1) \cdot \frac{37}{38} = 0$$

☺ Of course, the casino has to be able to make money over the long run, to be able to continue to take your money.

Example: Raffle

Example

A Raffle is going to sell 100 tickets for 8 prizes: one \$100 prize, two \$25 prizes and five \$10 prizes. What is a fair price for a ticket?

☞ Let X be the random variable for the expected winnings, and let x be the ticket price. If X is to be fair, then

$$\begin{aligned} E[X] &= 100 \cdot \frac{1}{100} + 25 \cdot \frac{2}{100} + 10 \cdot \frac{5}{100} - x \cdot \frac{92}{100} \\ &= 2 - x \cdot \frac{92}{100} \end{aligned}$$

A fair price x is when $E[X] = 0$, so

$$x = \frac{100}{92} \cdot 2 \approx 2.17$$

Sic Bo

Example

Sic Bo is an ancient Chinese dice game played with 3 dice. Various bets are possible at different payoffs. Here are some examples

- Big: score between 11 and 17 (inclusive) except triples,
- Small: score between 4 and 10 (inclusive) except triples,
- Specific Doubles: a specific double (ex. 2 ones)
- Specific Triple (or 'Alls'): specific triple
- Triples (or All 'Alls'): any triple

There are many other possible bets. (Chuck-a-luck is a simple variation of Sic Bo.)

Outcomes

k	Canonical outcomes						Total
3	(1,1,1)						1
4	(1,1,2)						3
5	(1,1,3)	(1,2,2)					6
6	(1,1,4)	(1,2,3)	(2,2,2)				10
7	(1,1,5)	(1,2,4)	(1,3,3)	(2,2,3)			15
8	(1,1,1)	(1,2,5)	(1,3,4)	(2,2,4)	(2,3,3)		21
9	(1,2,6)	(1,3,5)	(1,4,4)	(2,2,5)	(2,3,4)	(3,3,3)	25
10	(1,3,6)	(1,4,5)	(2,2,6)	(2,3,6)	(2,4,4)	(3,3,4)	27
11	(1,4,6)	(1,5,5)	(2,3,6)	(2,4,5)	(3,3,5)	(3,4,4)	27
12	(1,5,6)	(2,4,6)	(2,5,5)	(3,3,6)	(3,4,5)	(4,4,4)	25
13	(1,6,6)	(2,5,6)	(3,4,6)	(3,5,5)	(4,5,5)		21
14	(2,6,6)	(3,5,6)	(4,4,6)	(4,5,5)			15
15	(3,6,6)	(4,5,6)	(5,5,5)				10
16	(4,6,6)	(5,5,6)					6
17	(5,6,6)						3
18	(6,6,6)						1

Sic Bo

☞ Let X be the random variable of the payoff for various bets. Here is the expected value (at Atlantic City odds):

Game	Odds	Payoff	$E[X]$
Big	$\frac{105}{216}$	1	-0.028
Small	$\frac{105}{216}$	1	-0.028
Doubles	$\frac{16}{216}$	10	-0.185
'Alls'	$\frac{1}{216}$	180	-0.162
All 'Alls'	$\frac{6}{216}$	30	-0.139