Math 425
Introduction to Probability
Lecture 32

Kenneth Harris
kaharri@umich.edu

Department of Mathematics
University of Michigan

April 6, 2009

Expectation of Sums

☞ The expectation of a sum is a sum of expectations.

Theorem

If X and Y are random variables whose expectations exist and c, d constants, then

$$E[cX + dY] = cE[X] + dE[Y].$$

If $X_1, \ldots, X_n$ are random variables whose expectations exist and $c_1, \ldots, c_n$ constants, then

$$E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i E[X_i].$$

Indicator variables

☞ A random variable I is an indicator for the event A if

$$I(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \in A^c \end{cases}$$

Key Property. Let I be an indicator variable for E. Then

$$\begin{aligned} E[I] &= 0 \cdot \mathbf{P}\{A^c\} + 1 \cdot \mathbf{P}\{A\} \\ &= \mathbf{P}\{A\}. \end{aligned}$$

Use of Indicator variables

Use of Indicators. Suppose X is the number of events that occur among some collection of events $A_1, \ldots, A_n$. Let $X_1, \ldots, X_n$ be indicator variables for these events.

Then $X = \sum_{k=1}^{n} X_k$, so

$$E[X] = E[\sum_{k=1}^{n} X_k] = \sum_{i=1}^{k} E[X_k] = \sum_{i=1}^{k} \mathbf{P}\{A_k\}$$

When to use. The method of indicator variables is useful when we need to find $E[X]$ for a counting variable X, especially when we can break it down into counting the occurrences of events in a collection $A_1, \ldots, A_n$, where the probability $\mathbf{P}\{A_k\}$ is easy to compute.

Example: Hypergeometric mean

Example. Suppose n balls are drawn at random from an urn without replacement. There are r red balls and b blue balls where $n < r$ and $n < b$. What is the expected number of red balls drawn.

Solution. Let R count the number of red balls drawn, and R_i ($i \leq n$) be the indicator that the ith ball drawn is red:

$$R_i = \begin{cases} 1 & \text{if } i\text{th drawn ball is red,} \\ 0 & \text{otherwise.} \end{cases}$$

So, $R = \sum_{i=1}^{n} R_i$.

☞ The problem is now reduced to computing $E[R_i]$.

Example – continued

☞ The ith ball drawn is equally likely to be any of the $r + b$ balls, so

$$E[R_i] = \frac{r}{r+b}$$

☞ So, the expected number of red balls drawn is

$$E[R] = \sum_{i=1}^{n} E[R_i] = n \cdot \frac{r}{r+b}$$

This is exactly what we found in Chapter 4 (section 4.8.3).

Example: Negative Hypergeometric Mean

Example. Suppose we have an urn with r red balls and b blue balls. We are going to draw balls from the urn without replacement.

What is the expected number of draws required before k red balls are drawn?(Of course, $k \leq r$.)

☞ Let X count the number of balls drawn before there are k red balls.

The probability mass function for X (for $k \leq n \leq r + b$)

$$\mathbf{P}\{X = n\} = \frac{\binom{r}{k} \cdot \binom{b}{n-k}}{\binom{r+b}{n}}$$

X is said to be a negative hypergeometric random variable.

Example – continued

☞ Let the blue balls be numbered $x_1, \ldots, x_b$ and consider the indicator variables (for $i = 1, \ldots, b$)

$$X_i = \begin{cases} 1 & \text{if ball } b_i \text{ is chosen before } k \text{ red balls} \\ 0 & \text{otherwise.} \end{cases}$$

☞ Let X be the random variable counting the number balls drawn until k red balls are taken. So,

$$X = k + \sum_{i=1}^{b} X_i.$$

Then

$$E[X] = k + \sum_{i=1}^{b} E[X_i] = k + b \cdot E[X_1].$$

Reason. Since each ball is equally likely to be drawn the expectations $E[X_i]$ are the same for each i.

Example – continued

☞ Compute $\mathbf{P}\{X_1 = 1\}$. This is the probability of drawing the ball b_1 before k red balls are drawn.

Method 1. Ignoring order of draw.

$$E[X_1] = \mathbf{P}\{X_1 = 1\} = \frac{\binom{r}{k-1} \cdot \binom{1}{1}}{\binom{r+1}{k}} = \frac{k}{r+1}.$$

Reason. Choose k balls among the rred ones and b_1; we want only those ways which include b_1.

Method 2. Tracking order of draw.

$$E[X_1] = \mathbf{P}\{X_1 = 1\} = \sum_{i=1}^{k} \frac{1}{r+1} = \frac{k}{r+1}.$$

Reason. If b_1 is chosen before k red balls, then it was chosen in some position $i = 1, \ldots, k$ among $r+1$ possible balls.

Example – continued

☞ Compute the expected number of balls to be drawn until k red balls:

$$\begin{aligned} E[X] &= k + \sum_{i=1}^{b} E[X_i] \\ &= k + b \cdot E[X_1] \\ &= k + b \cdot \frac{k}{r+1} \\ &= \frac{k(r+b+1)}{r+1} \end{aligned}$$

Example: Counting Cards

Example. What is the expected number of cards draws before all 13 ♥ appear?

Solution. This is a problem involves the negative hypergeometric distribution where

- The red balls are ♥. So $r = 13$.
- The blue balls are the other three suits ♣, ♦, ♠. So $b = 39$.
- We want to draw $k = 13$ ♥.

Let X count the number of cards draw before all 13 ♥.

$$\begin{aligned} E[X] &= \frac{k(r+b+1)}{r+1} \\ &= \frac{13(13+39+1)}{13+1} = \frac{689}{14} \approx 49.214. \end{aligned}$$

Example: Screeing

Example

The concept of screening pooled samples originated during the Second World War to detect syphilis in U.S. soldiers. The strategy is used for testing a large number of samples (typically blood).

- Subgroups of at most 15 samples are chosen.
- Part of each specimen from each subgroup is pooled and tested.
- If a subgroup tests negative, then all of the individual samples in that subgroup are declared negative.
- If a subgroup tests positive, then each constituent sample of the subgroup is subsequently tested individually.

The use of pooling has been more extensively used recently in HIV testing of large specimen pools of low risk populations, such as in blood banks.

Example – continued

☞ 1000 soldiers are to be tested for syphillis. They will be broken into subgroups of 10. The samples in each subgroup will be tested, and if negative, the entire subgroup will be declared negative. However, if positive, each person in the group will be tested individually.

Suppose 1% of the soldier population has syphillis.

Question. How many tests can be expected to be performed using pooling, compared with testing each specimen individually?

☞ Testing every specimen individually requires 1000 tests.

Example – continued

☞ Let X_i ($i = 1, \ldots, 100$) be random variables which count the number of tests administered to the ith subgroup.

If some sample is positive, then 11 tests are performed, and otherwise only 1 test is performed. So, for each i

$$E[X_i] = 1 \cdot 0.99^{10} + 11 \cdot (1 - 0.99^{10}) \approx 1.956.$$

☞ Since each random variable is identically distributed, they have the same expectation. The number of tests expected is

$$E[X_1 + \ldots + X_{100}] \approx 100 \cdot 1.956 = 195.6.$$

By testing every individual we would require 1000 tests.

Example: Coupon Collecting Problem

Example

Each packet of some ineffably dull and noxious product contains one of N different types of flashy coupon. Each packet is equally likely to contain any of N types. How many packets should you expect to purchase until you first possess all N types?

Example – continued

☞ Let T count the number of packets you must purchase until you first possess all N types.

Break the problem down into simpler components using the counting variables (for $i = 1, 2, \ldots, N-1$)

- Let T_i be the number of further packets purchased after the $i-1$st new coupon until the ith new coupon.

Note: $T_1 = 1$, since any coupon you get will be new (you started with 0).

☞ T_i is a geometric random variable, where the probability of success (you collect a new coupon) is $p_i = \frac{N-i+1}{N}$ (there are $i-1$ out of N that are failures). So,

$$E[T_i] = \frac{1}{p_i} = \frac{N}{N-i+1}.$$

Example – continued

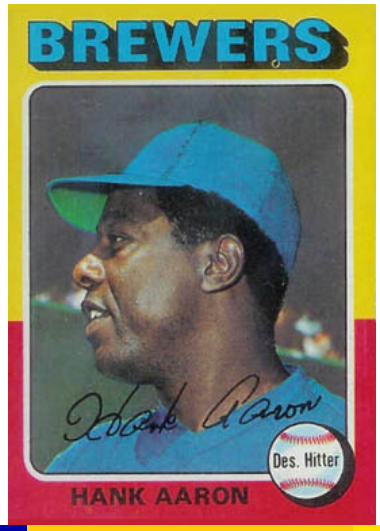
☞ Since $T = \sum_{i=1}^{N} T_i$, the expected number of packets is

$$\begin{aligned} E[T] &= \sum_{i=1}^{N} E[T_i] \\ &= \sum_{i=1}^{N} \frac{N}{N-i+1} \\ &= N\Big(\sum_{i=1}^{N} \frac{1}{i}\Big) \\ &\approx N \cdot \int_1^N \frac{1}{x}\, dx \\ &\approx N \log N + N\gamma \qquad \text{where } \gamma \approx 0.5772 \end{aligned}$$

γ is known as the Euler-Masheroni constant. (It is unknown whether it is rational or not.)

Example – continued

☞ As a kid I desparately needed Topps #660, Hank Aaron, to complete my collection:



Example – continued

☞ There were 660 cards in the complete set of 1975 Topps Baseball Series. I think there were 15 cards (and one ineffably dull and noxious product) in a packet, which was about 15 cents.

I should expect to have to collect alot of cards before I collected each card:

$$E[T] \approx 660 \ln 660 + (0.5772)660 \approx 4666 \text{ cards.}$$

☞ $46.66 is not chump change to a kid on a quarter a week allowance!!

Tail Sums

☞ There is usually more than one way for a counting variable X to be expressed as the sum of indicator variables for a collection of events.

To find $E[X]$ you need onlyone such collection.

☞ When $0 \le X \le n$, we can write $X = \sum_{i=1}^{n} X_i$ where

$$X_i = \begin{cases} 1 & \text{if } X \ge i \\ 0 & \text{if } X < i \end{cases} \qquad 1 \le i \le n.$$

Thus,

$$E[X] = \sum_{i=1}^{n} E[X_i].$$

Tail Sums

Theorem

For X a discrete random variable with possible values $\{0, 1, 2, \ldots, n\}$,

$$E[X] = \sum_{i=1}^{n} \boldsymbol{P}\{X \geq i\}$$

Proof.

Let X_i be the indicator variable for $\mathbf{P}\{X \geq i\}$.

If $X = k > 0$, then

$$X_1 = X_2 = \ldots = X_k = 1 \quad \text{and} \quad X_{k+1} = X_{k+2} \ldots = X_n = 0.$$

If $X = 0$, then $X_1 = X_2 = \ldots = X_n = 0$. Either way $X = \sum_{i=1}^{n} X_i$. □

Example

Example. Throw four dice. Let M be the minimum value. FInd $E[M]$.

Solution. For any $1 \leq j \leq 6$, the event $\{M \geq j\}$ means that each of $X_1, \ldots, X_4 \geq j$ (where X_i is the value of the ith die).

$$\mathbf{P}\{M \geq j\} \;=\; \mathbf{P}\{X_1 \geq j, X_2 \geq j, X_3 \geq j, X_4 \geq j\} \;=\; \left(\frac{6-j+1}{6}\right)^4$$

by the independence of $X_1, \ldots, X_4$.

☞ The expected minimal value is

$$\begin{aligned} E[M] &= \sum_{j=1}^{6} \mathbf{P}\{M \geq j\} \\ &= \left(\frac{1}{6}\right)^4 + \left(\frac{2}{6}\right)^4 + \left(\frac{3}{6}\right)^4 + \left(\frac{4}{6}\right)^4 + \left(\frac{5}{6}\right)^4 + \left(\frac{6}{6}\right)^4 \\ &\approx 1.755. \end{aligned}$$

Example

Example. Throw four die. Let S be the sum of the largest three of the values. Find $E[S]$.

Solution. Note that $S = T - M$ where T is the sum of all four die and M is the minimum value.

Since $T = X_1 + X_2 + X_3 + X_4$ (where X_i is the value of the ith die)

$$E[T] \;=\; \sum_{i=1}^{4} E[X_i] \;=\; 4 \cdot 3.5 \;=\; 14.$$

So,

$$E[S] \;=\; E[T - M] \;=\; E[T] - E[M] \;\approx\; 14 - 1.755 \;=\; 12.245.$$

Example

Example

A 52 card deck is randomly shuffled so that all 52! orderings are equally likely. You are going to make 52 guesses sequentially, where the ith guess is on the card in position ith.

Suppose after each guess you are shown the card that was in the position. What is the expected number of correct guesses with the following strategy

- Guess at random a card on the ith draw which has not appeared in the first $i - 1$ draws.

This strategy maximizes the expected number of correct guesses.

Example

☞ Let X count the number of correct guesses, and X_i $(i = 1, \ldots, 52)$ the indicator variables

$$X_i = \begin{cases} 1 & \text{if the } i\text{th card is correctly guessed} \\ 0 & \text{otherwise.} \end{cases}$$

So $X = \sum_{i=1}^{52} X_i$.

☞ Since $E[X_i] = \frac{1}{52-i+1}$,

$$\begin{aligned} E[X] &= \sum_{i=1}^{52} E[X_i] \\ &= \sum_{i=1}^{52} \frac{1}{52-i+1} = \frac{1}{52} + \frac{1}{51} + \ldots + \frac{1}{2} + 1 \\ &\approx \int_1^{52} \frac{1}{x}\, dx \;=\; \log 52 \approx 3.951 \end{aligned}$$

Example

Example

A 52 card deck is randomly shuffled so that all 52! orderings are equally likely. You are going to make 52 guesses sequentially, where the ith guess is on the card in position ith.

Suppose you are not shown the card after each guess, but only told whether you were right or wrong. What is the expected number of correct guesses with the following strategy

- Choose an ordering of 52 cards, and continue to guess the ith card (starting with $i = 1$) in the ordering until it appears (or the cards run out). Should this card appear, start guessing the $i + 1$st card.

This strategy maximizes the expected number of correct guesses.

Example

☞ Let X count the number of correct guesses and use the tail sums $X_k = \mathbf{P}\{X \geq k\}$ for $k = 1, \ldots, 52$.

The only way for $X \geq k$ is for the first k cards you will guess to appear somewhere in the deck in the order you guess. So,

$$\mathbf{P}\{X \geq k\} = \frac{1}{k!}$$

☞ Since $X = \sum_{k=1}^{52} X_k$

$$\begin{aligned} E[X] &= \sum_{k=1}^{52} E[X_k] \\ &= \sum_{k=1}^{52} \frac{1}{k!} = 1 + \frac{1}{2} + \frac{1}{3!} + \ldots + \frac{1}{52!} \\ &\approx e - 1 \approx 1.718 \end{aligned}$$

Note: $e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!} + \ldots$.