Averaging and Expectation

Suppose you toss a coin that's biased towards heads (Pr(heads) = 2/3) twice. How many heads do you expect to get?

- In mathematics-speak:
  What's the *expected number* of heads?

What about if you toss the coin k times?

What's the average weight of the people in this classroom?

- That's easy: add the weights and divide by the number of people in the class.

But what about if I tell you I'm going to toss a coin to determine which person in the class I'm going to choose; if it lands heads, I'll choose someone at random from the first aisle, and otherwise I'll choose someone at random from the last aisle.

- What's the expected weight?

Averaging makes sense if you use an equiprobable distribution; in general, we need to talk about *expectation*.

Random Variables

To deal with expectation, we formally associate with every element of a sample space a real number.

Definition: A *random variable* on sample space S is a function from S to the real numbers.

Example: Suppose we toss a biased coin ($\Pr(h) = 2/3$) twice. The sample space is:

- hh - Probability 4/9
- ht - Probability 2/9
- th - Probability 2/9
- tt - Probability 1/9

If we're interested in the number of heads, we would consider a random variable $\#H$ that counts the number of heads in each sequence:

$$\#H(hh) = 2;\quad \#H(ht) = \#H(th) = 1;\quad \#H(tt) = 0$$

Example: If we're interested in weights of people in the class, the sample space is people in the class, and we could have a random variable that associates with each person his or her weight.

Probability Distributions

If X is a random variable on sample space S, then the probability that X takes on the value c is

$$\Pr(X = c) = \Pr(\{s \in S \mid X(s) = c\})$$

Similarly,

$$\Pr(X \leq c) = \Pr(\{s \in S \mid X(s) \leq c\}.$$

This makes sense since the range of X is the real numbers.

Example: In the coin example,

$$\Pr(\#H = 2) = 4/9 \text{ and } \Pr(\#H \leq 1) = 5/9$$

Given a probability measure Pr on a sample space S and a random variable X, the *probability distribution* associated with X is $f_X(x) = \Pr(X = x)$.

- f_X is a probability measure on the real numbers.

The *cumulative distribution* associated with X is $F_X(x) = \Pr(X \leq x)$.

An Example With Dice

Suppose S is the sample space corresponding to tossing a pair of fair dice: $\{(i, j) \mid 1 \leq i, j \leq 6\}$.

Let X be the random variable that gives the sum:

- $X(i, j) = i + j$

$$\begin{aligned}
&f_X(2) = \Pr(X = 2) = \Pr(\{(1, 1)\}) = 1/36\\
&f_X(3) = \Pr(X = 3) = \Pr(\{(1, 2), (2, 1)\}) = 2/36\\
&\vdots\\
&f_X(7) = \Pr(X = 7) = \Pr(\{(1, 6), (2, 5), \ldots, (6, 1)\}) = 6/36\\
&\vdots\\
&f_X(12) = \Pr(X = 12) = \Pr(\{(6, 6)\}) = 1/36
\end{aligned}$$

Can similarly compute the cumulative distribution:

$$\begin{aligned}
&F_X(2) = f_X(2) = 1/36\\
&F_X(3) = f_X(2) + f_X(3) = 3/36\\
&\vdots\\
&F_X(12) = 1
\end{aligned}$$

The Finite Uniform Distribution

The finite uniform distribution is an equiprobable distribution. If $S = \{x_1, \ldots, x_n\}$, where $x_1 < x_2 < \ldots < x_n$, then:

$$f(x_k) = 1/n$$
$$F(x_k) = k/n$$

The Binomial Distribution

Suppose there is an experiment with probability p of success and thus probability $q = 1 - p$ of failure.

- For example, consider tossing a biased coin, where $\Pr(h) = p$. Getting "heads" is success, and getting tails is failure.

Suppose the experiment is repeated independently n times.

- For example, the coin is tossed n times.

This is called a sequence of *Bernoulli trials*.

Key features:

- Only two possibilities: success or failure.
- Probability of success does not change from trial to trial.
- The trials are independent.

What is the probability of k successes in n trials?

Suppose $n = 5$ and $k = 3$. How many sequences of 5 coin tosses have exactly three heads?

- *hhhtt*
- *hhtht*
- *hhtth*

⋮

$C(5, 3)$ such sequences!

What is the probability of each one?

$$p^3(1-p)^2$$

Therefore, probability is $C(5, 3)p^3(1-p)^2$.

Let $B_{n,p}(k)$ be the probability of getting k successes in n Bernoulli trials with probability p of success.

$$B_{n,p}(k) = C(n, k)p^k(1-p)^{n-k}$$

Not surprisingly, $B_{n,p}$ is called the *Binomal Distribution*.

The Poisson Distribution

A large call center receives, on average, λ calls/minute.

- What is the probability that exactly k calls come during a given minute?

Understanding this probability is critical for staffing!

- Similar issues arise if a printer receives, on average λ jobs/minute, a site gets λ hits/minute, ...

This is modelled well by the *Poisson distribution* with parameter λ:

$$f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- $f_\lambda(0) = e^{-\lambda}$
- $f_\lambda(1) = e^{-\lambda}\lambda$
- $f_\lambda(2) = e^{-\lambda}\lambda^2/2$

$e^{-\lambda}$ is a normalization constant, since

$$1 + \lambda + \lambda^2/2 + \lambda^3/3! + \cdots = e^\lambda$$

Deriving the Poisson

Poisson distribution = limit of binomial distributions.

Suppose at most one call arrives in each second.

- Since λ calls come each minute, expect about $\lambda/60$ each second.
- The probability that k calls come is $B_{60,\lambda/60}(k)$

This model doesn't allow more than one call/second. What's so special about 60? Suppose we divide one minute into n time segments.

- Probability of getting a call in each segment is λ/n.
- Probability of getting k calls in a minute is

$$\begin{array}{rl} & B_{n,\lambda/n}(k) \\ = & C(n,k)(\lambda/n)^k(1-\frac{\lambda}{n})^{n-k} \\ = & C(n,k)\left(\frac{\lambda/n}{1-\frac{\lambda}{n}}\right)^k\left(1-\frac{\lambda}{n}\right)^n \\ = & \frac{\lambda^k}{k!}\frac{n!}{(n-k)!}\left(\frac{1}{n-\lambda}\right)^k\left(1-\frac{\lambda}{n}\right)^n \end{array}$$

Now let $n \rightarrow \infty$:

- $\lim_{n\rightarrow\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$
- $\lim_{n\rightarrow\infty}\frac{n!}{(n-k)!}\left(\frac{1}{n-\lambda}\right)^k = 1$

Conclusion: $\lim_{n\rightarrow\infty} B_{n,\lambda/n}(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

New Distributions from Old

If X and Y are random variables on a sample space S, so is $X+Y$, $X+2Y$, XY, $\sin(X)$, etc.

For example,

- $(X+Y)(s) = X(s)+Y(s)$.
- $\sin(X)(s) = \sin(X(s))$

Note $\sin(X)$ is a random variable: a function from the sample space to the reals.

Some Examples

Example 1: A fair die is rolled. Let X denote the number that shows up. What is the probability distribution of $Y = X^2$?

$$\begin{array}{rl} \{s : Y(s) = k\} & = \{s : X^2(s) = k\} \\ & = \{s : X(s) = -\sqrt{k}\} \cup \{s : X(s) = \sqrt{k}\}. \end{array}$$

Conclusion: $f_Y(k) = f_X(\sqrt{k}) + f_X(-\sqrt{k})$.
So $f_Y(1) = f_Y(4) = f_Y(9) = \cdots f_Y(36) = 1/6$.
$f_Y(k) = 0$ if $k \notin \{1,4,9,16,25,36\}$.

Example 2: A coin is flipped. Let X be 1 if the coin shows H and -1 if T. Let $Y = X^2$.

- In this case $Y \equiv 1$, so $\Pr(Y = 1) = 1$.

Example 3: If two dice are rolled, let X be the number that comes up on the first dice, and Y the number that comes up on the second.

- Formally, $X((i,j)) = i$, $Y((i,j)) = j$.

The random variable $X+Y$ is the total number showing.

Example 4: Suppose we toss a biased coin n times (more generally, we perform n Bernoulli trials). Let X_k describe the outcome of the kth coin toss: $X_k = 1$ if the kth coin toss is heads, and 0 otherwise.

How do we formalize this?

- What's the sample space?

Notice that $\Sigma_{k=1}^n X_k$ describes the number of successes of n Bernoulli trials.

- If the probability of a single success is p, then $\Sigma_{k=1}^n X_k$ has distribution $B_{n,p}$
  - The binomial distribution is the sum of Bernoullis

Independent random variables

In a roll of two dice, let X and Y record the numbers on the first and second die respectively.

- What can you say about the events $X = 3$, $Y = 2$?
- What about $X = i$ and $Y = j$?

Definition: The random variables X and Y are independent if for every x and y the events $X = x$ and $Y = y$ are independent.

Example: X and Y above are independent.

Definition: The random variables $X_1, X_2, \ldots, X_n$ are *mutually independent* if, for every $x_1, x_2 \ldots, x_n$

$\Pr(X_1 = x_1 \cap \ldots \cap X_n = x_n) = \Pr(X_1 = x_1) \ldots \Pr(X_n = x_n)$

Example: X_k, the success indicators in n Bernoulli trials, are independent.

Pairwise vs. mutual independence

Mutual independence implies pairwise independence; the converse may not be true:

Example 1: A ball is randomly drawn from an urn containing 4 balls: one blue, one red, one green and one multicolored (red + blue + green)

- Let X_1, X_2 and X_3 denote the indicators of the events the ball has (some) blue, red and green respectively.
- $\Pr(X_i = 1) = 1/2$, for $i = 1, 2, 3$

X_1 and X_2 independent:

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	1/4	1/4
$X_2 = 1$	1/4	1/4

Similarly, X_1 and X_3 are independent; so are X_2 and X_3.

Are X_1, X_2 and X_3 independent? No!

$$\Pr(X_1 = 1 \cap X_2 = 1 \cap X_3 = 1) = 1/4$$
$$\Pr(X_1 = 1)\Pr(X_2 = 1)\Pr(X_3 = 1) = 1/8.$$

Example 2: Suppose X_1 and X_2 are bits (0 or 1) chosen uniformly at random; $X_3 = X_1 \oplus X_2$.

- X_1, X_2 are independent, as are X_1, X_3 and X_2, X_3
- But X_1, X_2, and X_3 are not mutually independent
  - X_1 and X_2 together determine X_3!

The distribution of $X + Y$

Suppose X and Y are independent random variables whose range is included in $\{0, 1, \ldots, n\}$. For $k \in \{0, 1, \ldots, 2n\}$,

$$(X + Y = k) = \cup_{j=0}^{k} \left((X = j) \cap (Y = k - j)\right).$$

Note that some of the events might be empty

- E.g., $X = k$ is bound to be empty if $k > n$.

This is a disjoint union so

$$\begin{aligned} & \Pr(X + Y = k) \\ = \; & \Sigma_{j=0}^{k} \Pr(X = j \cap Y = k - j) \\ = \; & \Sigma_{j=0}^{k} \Pr(X = j)\Pr(Y = k - j) \quad \text{[by independence]} \end{aligned}$$

Example: The Sum of Binomials

Suppose X has distribution $B_{n,p}$, Y has distribution $B_{m,p}$, and X and Y are independent.

$$\begin{aligned} & \Pr(X + Y = k) \\ = \; & \Sigma_{j=0}^{k} \Pr(X = j \cap Y = k - j) \quad \text{[sum rule]} \\ = \; & \Sigma_{j=0}^{k} \Pr(X = j)\Pr(Y = k - j) \quad \text{[independence]} \\ = \; & \Sigma_{j=0}^{k} \tbinom{n}{j} p^j (1-p)^{n-j} \tbinom{m}{k-j} p^{k-j} (1-p)^{m-k+j} \\ = \; & \Sigma_{j=0}^{k} \tbinom{n}{j} \tbinom{m}{k-j} p^k (1-p)^{n+m-k} \\ = \; & (\Sigma_{j=0}^{k} \tbinom{n}{j} \tbinom{m}{k-j}) p^k (1-p)^{n+m-k} \\ = \; & \tbinom{n+m}{k} p^k (1-p)^{n+m-k} \\ = \; & B_{n+m,p}(k) \end{aligned}$$

Thus, $X + Y$ has distribution $B_{n+m,p}$.

An easier argument: Perform $n + m$ Bernoulli trials. Let X be the number of successes in the first n and let Y be the number of successes in the last m. X has distribution $B_{n,p}$, Y has distribution $B_{m,p}$, X and Y are independent, and $X + Y$ is the number of successes in all $n + m$ trials, and so has distribution $B_{n+m,p}$.

Expected Value

Suppose we toss a biased coin, with $\Pr(h) = 2/3$. If the coin lands heads, you get \$1; if the coin lands tails, you get \$3. What are your expected winnings?

- 2/3 of the time you get \$1;
  1/3 of the time you get \$3
- $(2/3 \times 1) + (1/3 \times 3) = 5/3$

What's a good way to think about this? We have a random variable W (for winnings):

- $W(h) = 1$
- $W(t) = 3$

The expectation of W is

$$\begin{aligned} E(W) &= \Pr(h)W(h) + \Pr(t)W(t) \\ &= \Pr(W = 1) \times 1 + \Pr(W = 3) \times 3 \end{aligned}$$

More generally, the *expected value* of random variable X on sample space S is

$$E(X) = \Sigma_x x \Pr(X = x)$$

An equivalent definition:

$$E(X) = \Sigma_{s \in S} X(s) \Pr(s)$$

Example: What is the expected count when two dice are rolled?

Let X be the count (the sum of the values on the two dice):

$$\begin{aligned} & E(X) \\ = \ & \Sigma_{i=2}^{12} i \Pr(X = i) \\ = \ & 2\tfrac{1}{36} + 3\tfrac{2}{36} + 4\tfrac{3}{36} + \cdots + 7\tfrac{6}{36} + \cdots + 12\tfrac{1}{36} \\ = \ & \tfrac{252}{36} \\ = \ & 7 \end{aligned}$$

Expectation of Binomials

What is $E(B_{n,p})$, the expectation for the binomial distribution $B_{n,p}$

- How many heads do you expect to get after n tosses of a biased coin with $\Pr(h) = p$?

Method 1: Use the definition and crank it out:

$$E(B_{n,p}) = \Sigma_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

This looks awful, but it can be calculated ...

Method 2: Use Induction; break it up into what happens on the first toss and on the later tosses.

- On the first toss you get heads with probability p and tails with probability $1 - p$. On the last $n - 1$ tosses, you expect $E(B_{n-1,p})$ heads. Thus, the expected number of heads is:

$$\begin{aligned} E(B_{n,p}) &= p(1 + E(B_{n-1,p})) + (1-p)(E(B_{n-1,p})) \\ &= p + E(B_{n-1,p}) \\ E(B_{1,p}) &= p \end{aligned}$$

Now an easy induction shows that $E(B_{n,p}) = np$.

There's an even easier way . . .

Expectation is Linear

Theorem: $E(X + Y) = E(X) + E(Y)$

Proof: Recall that

$$E(X) = \Sigma_{s \in S} \Pr(s)X(s)$$

Thus,

$$\begin{aligned} E(X + Y) &= \Sigma_{s \in S} \Pr(s)(X + Y)(s) \\ &= \Sigma_{s \in S} \Pr(s)X(s) + \Sigma_{s \in S} \Pr(s)Y(s) \\ &= E(X) + E(Y). \end{aligned}$$

Theorem: $E(aX) = aE(X)$

Proof:

$$E(aX) = \Sigma_{s \in S} \Pr(s)(aX)(s) = a\Sigma_{s \in S} X(s) = aE(X).$$

Example 1: Back to the expected value of tossing two dice:
Let X_1 be the count on the first die, X_2 the count on the second die, and let X be the total count.

Notice that

$$E(X_1) = E(X_2) = (1+2+3+4+5+6)/6 = 3.5$$

$$E(X) = E(X_1+X_2) = E(X_1)+E(X_2) = 3.5+3.5 = 7$$

Example 2: Back to the expected value of $B_{n,p}$.

Let X be the total number of successes and let X_k be the outcome of the kth experiment, $k = 1, \ldots, n$:

$$E(X_k) = p \cdot 1 + (1-p) \cdot 0 = p$$

$$X = X_1 + \cdots + X_n$$

Therefore

$$E(X) = E(X_1) + \cdots + E(X_n) = np.$$

Expectation of Poisson Distribution

Let X be Poisson with parameter λ: $f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k \in N$.

$$\begin{aligned} E(X) &= \Sigma_{k=0}^{\infty} k \cdot e^{-\lambda}\tfrac{\lambda^k}{k!} \\ &= \Sigma_{k=1}^{\infty} e^{-\lambda}\tfrac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda}\Sigma_{k=1}^{\infty}\tfrac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda}\Sigma_{j=0}^{\infty}\tfrac{\lambda^j}{j!} \\ &= \lambda e^{-\lambda}e^{\lambda} \qquad \text{[Taylor series!]} \\ &= \lambda \end{aligned}$$

Does this make sense?

- Recall that, for example, X models the number of incoming calls for a tech support center whose average rate per minute is λ.

Geometric Distribution

Consider a sequence of Bernoulli trials. Let X denote the number of the first successful trial.

- E.g., the first time you see heads

X has a *geometric* distribution.

$$f_X(k) = (1-p)^{k-1}p \qquad k \in N^+.$$

- The probability of seeing heads for the first time on the kth toss is the probability of getting $k-1$ tails followed by heads
- This is also called a *negative binomial* distribution of order 1.
  - The negative binomial of order n gives the probability that it will take k trials to have n successes