Conditional Expectation

- X and Y are jointly discrete random variables
  - Recall conditional PMF of X given Y = y:

$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

- Define conditional expectation of X given Y = y:

$$E[X \mid Y = y] = \sum_x x P(X = x \mid Y = y) = \sum_x x\, p_{X|Y}(x \mid y)$$

- Analogously, jointly continuous random variables:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad E[X \mid Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx$$

Rolling Dice

- Roll two 6-sided dice D_1 and D_2
  - X = value of $D_1 + D_2$ Y = value of D_2
  - What is E[X | Y = 6]?

$$E[X \mid Y = 6] = \sum_x x P(X = x \mid Y = 6)$$

$$= \left(\frac{1}{6}\right)(7 + 8 + 9 + 10 + 11 + 12) = \frac{57}{6} = 9.5$$

  - Intuitively makes sense: 6 + E[value of D_1] = 6 + 3.5

Hyper for the Hypergeometric

- X and Y are independent random variables
  - X ~ Bin(n, p)   Y ~ Bin(n, p)
  - What is E[X | X + Y = m], where m ≤ n?
  - Start by computing P(X = k | X + Y = m):

$$P(X = k \mid X + Y = m) = \frac{P(X = k, X + Y = m)}{P(X + Y = m)} = \frac{P(X = k, Y = m - k)}{P(X + Y = m)} = \frac{P(X = k)P(Y = m - k)}{P(X + Y = m)}$$

$$= \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot \binom{n}{m-k} p^{m-k} (1-p)^{n-(m-k)}}{\binom{2n}{m} p^m (1-p)^{2n-m}} = \frac{\binom{n}{k} \cdot \binom{n}{m-k}}{\binom{2n}{m}}$$

  - Hypergeometric: (X | X + Y = m) ~ HypG(*m*, *2n, n*)
  - E[X | X + Y = m] = *nm*/2*n* = *m*/2

total draws

total balls

white balls

Properties of Conditional Expectation

- X and Y are jointly distributed random variables

$$E[g(X)\,|\,Y=y]=\sum_{x} g(x)\,p_{X|Y}(x\,|\,y) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x)\,f_{X|Y}(x\,|\,y)\,dx$$

- Expectation of conditional sum:

$$E\left[\sum_{i=1}^{n} X_i \,|\, Y=y\right]=\sum_{i=1}^{n} E[X_i \,|\, Y=y]$$

Expectations of Conditional Expectations

- Define g(Y) = E[X | Y]
  - g(Y) is a random variable
  - For any Y = y, g(Y) = E[X | Y = y]
    - This is just function of Y, since we sum over all values of X
  - What is E[E[X | Y]] = E[g(Y)]? (Consider discrete case)

$$E[E[X \mid Y]] = \sum_y E[X \mid Y = y] P(Y = y)$$

$$= \sum_y [\sum_x x P(X = x \mid Y = y)] P(Y = y)$$

$$= \sum_y \sum_x x P(X = x, Y = y) = \sum_x x \sum_y P(X = x, Y = y)$$

$$= \sum_x x P(X = x) = E[X]$$

(Same for continuous)

Analyzing Recursive Code

```
int Recurse() {
    int x = randomInt(1, 3);  // Equally likely values

    if (x == 1) return 3;
    else if (x == 2) return (5 + Recurse());
    else return (7 + Recurse());
}
```

- Let Y = value returned by **`Recurse()`**. What is E[Y]?

$$E[Y] = E[Y \mid X = 1]P(X = 1) + E[Y \mid X = 2]P(X = 2) + E[Y \mid X = 3]P(X = 3)$$

$$E[Y \mid X = 1] = 3$$

$$E[Y \mid X = 2] = E[5 + Y] = 5 + E[Y]$$

$$E[Y \mid X = 3] = E[7 + Y] = 7 + E[Y]$$

$$E[Y] = 3(1/3) + (5 + E[Y])(1/3) + (7 + E[Y])(1/3) = (1/3)(15 + 2E[Y])$$

$$E[Y] = 15$$

Random Number of Random Variables

- Say you have a web site: **`PimentoLoaf.com`**
  - X = Number of people/day visit your site. X ~ N(50, 25)
  - Y_i = Number of minutes spent by visitor *i*. Y_i ~ Poi(8)
  - X and all Y_i are independent
  - Time spent by all visitors/day: $W = \sum_{i=1}^{X} Y_i$. What is E[W]?

$$E[W] = E\left[\sum_{i=1}^{X} Y_i\right] = E\left[E\left[\sum_{i=1}^{X} Y_i \mid X\right]\right] = E\big[X \cdot E[Y_i]\big] = E[X]E[Y_i] = 50 \cdot 8$$

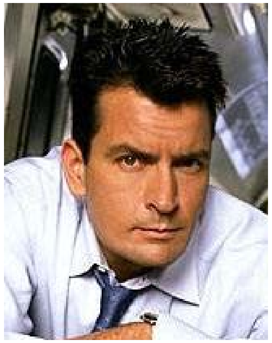
$$E\left[\sum_{i=1}^{X} Y_i \mid X = n\right] = \sum_{i=1}^{n} E[Y_i \mid X = n] = \sum_{i=1}^{n} E[Y_i] = nE[Y_i]$$

$$E\left[\sum_{i=1}^{X} Y_i \mid X\right] = X \cdot E[Y_i]$$

Making Predictions

- We observe random variable X
  - Want to make prediction about Y
  - E.g., X = stock price at 9am, Y = stock price at 10am
  - Let $g(X)$ be function we use to predict Y, i.e.: $\hat{Y} = g(X)$
  - Choose $g(X)$ to minimize E[(Y – $g(X)$)^2]
  - Best predictor: $g(X)$ = E[Y | X]
  - Intuitively: E[(Y – c)^2] is minimized when c = E[Y]
    - Now, you observe X, and Y depends on X, then use c = E[Y | X]
  - You just got your first baby steps into Machine Learning
    - We'll go into this more rigorously in a few weeks

Speaking of Babies...

- Say my height is X inches ($x = 71$)
  - My son:  He does not look like: 

Speaking of Babies...

- Say my height is *X* inches ($x = 71$)
  - My son:  But, perhaps a bit like: 
  - Say, historically, sons grow to heights $Y \sim N(X + 1, 4)$, where *X* is height of father
    - $Y = (X + 1) + C$ where $C \sim N(0, 4)$
  - What should I predict for the eventual height of my son?
  - $E[Y \mid X = 71] = E[X + 1 + C \mid X = 71]$
    $= E[72 + C] = E[72] + E[C] = 72 + 0$
    $= 72$ inches

Computing Probabilities by Conditioning

- X = indicator variable for event A: $X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$
  - E[X] = P(A)
  - Similarly, E[X | Y = y] = P(A | Y = y) for any Y
  - So: $E[X] = E_Y[\, E_X[X \mid Y]\,] = E[\, E[X \mid Y]\,] = E[\, P(A \mid Y)\,]$
  - In discrete case:

$$E[X] = \sum_y P(A \mid Y = y) P(Y = y) = P(A)$$

    - Also holds analogously in continuous case
  - Generalize, defining indicator variables $F_i = (Y = y_i)$:

$$P(A) = \sum_{i=1}^{n} P(A \mid F_i) P(F_i)$$

    - Called “Law of total probability”

Hiring Software Engineers

- Interviewing *n* software engineer candidates
  - All *n*! orderings equally likely, but only hiring 1 candidate
    - Claim: There is α-to-1 factor difference in productivity between the "best" and "average" software engineer
    - Steve Jobs set α = 25, Mark Zuckerberg claimed α = 100
  - Right after each interview must decide hire/no hire
  - Feedback from interview of candidate *i* is just relative ranking with respect to previous *i* – 1 candidates
    - Strategy: first interview *k* (of *n*) candidates, then hire next candidate better than all of first *k* candidates
    - P_k(best) = probability that best of all *n* candidates is hired
    - X = position of best candidate (1, 2, ..., *n*)

$$P_k(\text{Best}) = \sum_{i=1}^{n} P_k(\text{Best} \mid X = i) P(X = i) = \frac{1}{n} \sum_{i=1}^{n} P_k(\text{Best} \mid X = i)$$

Hiring Software Engineers (cont.)

- Note: $P_k(\text{Best} \mid X = i) = 0 \quad \text{if} \quad i \le k$
- We will select best candidate (in position *i*) if best of first *i* – 1 candidates is among the first *k* interviewed

$$P_k(\text{Best} \mid X = i) = P_k(\text{best of first } i-1 \text{ in first } k \mid X = i) = \frac{k}{i-1} \quad \text{if} \quad i > k$$

$$P_k(\text{Best}) = \frac{1}{n}\sum_{i=1}^{n} P_k(\text{Best} \mid X = i) = \frac{1}{n}\sum_{i=k+1}^{n}\frac{k}{i-1}$$

$$\approx \frac{k}{n}\int_{i=k+1}^{n}\frac{1}{i-1}\,di = \frac{k}{n}\ln(i-1)\Big|_{k+1}^{n} = \frac{k}{n}\ln\frac{n-1}{k} \approx \frac{k}{n}\ln\frac{n}{k}$$

- To maximize, differentiate P_k(Best) with respect to *k*:

$$g(k) = \frac{k}{n}\ln\frac{n}{k} \qquad g'(k) = \frac{1}{n}\ln\frac{n}{k} + \frac{k}{n}\left(\frac{-1}{k}\right) = \frac{1}{n}\ln\frac{n}{k} - \frac{1}{n}$$

- Set $g'(k) = 0$ and solve for *k*:

$$\frac{1}{n}\ln\frac{n}{k} - \frac{1}{n} = 0 \;\Rightarrow\; \ln\frac{n}{k} = 1 \;\Rightarrow\; \frac{n}{k} = e \;\Rightarrow\; k = \frac{n}{e}$$

- Interview n/e candidates, then pick best: P_k(Best) ≈ 1/e ≈ 0.368