Independent Discrete Variables

- Two discrete random variables X and Y are called **independent** if:

$$p(x,y) = p_X(x)p_Y(y) \quad \text{for all } x, y$$

- Intuitively: knowing the value of X tells us nothing about the distribution of Y (and vice versa)
  - If two variables are **not** independent, they are called ***dependent***
- Similar conceptually to independent *events*, but we are dealing with multiple ***variables***
  - Keep your events and variables distinct (and clear)!

Coin Flips

- Flip coin with probability *p* of "heads"
  - Flip coin a total of *n* + *m* times
  - Let X = number of heads in first *n* flips
  - Let Y = number of heads in next *m* flips

$$P(X = x, Y = y) = \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{y} p^y (1-p)^{m-y}$$
$$= P(X = x)P(Y = y)$$

  - X and Y are independent
  - Let Z = number of total heads in *n* + *m* flips
  - Are X and Z independent?
    - What if you are told Z = 0?

Web Server Requests

- Let N = # of requests to web server/day
  - Suppose N ~ Poi(λ)
  - Each request comes from a human (probability = *p*) or from a "bot" (probability = (1 – *p*)), independently
  - X = # requests from humans/day (X | N) ~ Bin(N, *p*)
  - Y = # requests from bots/day (Y | N) ~ Bin(N, 1 - *p*)

$$P(X = i, Y = j) = P(X = i, Y = j \mid X + Y = i + j)P(X + Y = i + j) + P(X = i, Y = j \mid X + Y \neq i + j)P(X + Y \neq i + j)$$

  - Note: $P(X = i, Y = j \mid X + Y \neq i + j) = 0$

$$P(X = i, Y = j \mid X + Y = i + j) = \binom{i+j}{i} p^i (1-p)^j$$
$$P(X + Y = i + j) = e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$
$$P(X = i, Y = j) = \binom{i+j}{i} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$

Web Server Requests (cont.)

- Let N = # of requests to web server/day
  - Suppose N ~ Poi(λ)
  - Each request comes from a human (probability = *p*) or from a "bot" (probability = (1 – *p*)), independently
  - X = # requests from humans/day X | N ~ Bin(N, *p*)
  - Y = # requests from bots/day Y | N ~ Bin(N, 1 - *p*)

$$P(X = i, Y = j) = \frac{(i+j)!}{i!\,j!} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!} = e^{-\lambda} \frac{(\lambda p)^i}{i!\,j!} (\lambda(1-p))^j$$
$$= e^{-\lambda p} \frac{(\lambda p)^i}{i!} \cdot e^{-\lambda(1-p)} \frac{(\lambda(1-p))^j}{j!} = P(X = i)P(Y = j)$$

  where X ~ Poi(λp) and Y ~ Poi(λ(1 – *p*))
  - X and Y are independent!

Independent Continuous Variables

- Two continuous random variables X and Y are called **independent** if:

  P(X ≤ *a*, Y ≤ *b*) = P(X ≤ *a*) P(Y ≤ *b*) for any *a*, *b*

- Equivalently:

$$F_{X,Y}(a,b) = F_X(a)F_Y(b) \quad \text{for all } a, b$$
$$f_{X,Y}(a,b) = f_X(a)f_Y(b) \quad \text{for all } a, b$$

- More generally, joint density factors separately:

$$f_{X,Y}(x,y) = h(x)g(y) \quad \text{where } -\infty < x, y < \infty$$

Pop Quiz (Just Kidding…)

- Consider joint density function of X and Y:

$$f_{X,Y}(x,y) = 6e^{-3x}e^{-2y} \quad \text{for } 0 < x, y < \infty$$

  - Are X and Y independent? Yes!

Let $h(x) = 3e^{-3x}$ and $g(y) = 2e^{-2y}$, so $f_{X,Y}(x,y) = h(x)g(y)$

- Consider joint density function of X and Y:

$$f_{X,Y}(x,y) = 4xy \quad \text{for } 0 < x, y < 1$$

  - Are X and Y independent? Yes!

  Let $h(x) = 2x$ and $g(y) = 2y$, so $f_{X,Y}(x,y) = h(x)g(y)$
  - Now add constraint that: $0 < (x + y) < 1$
  - Are X and Y independent? No!
    - Cannot capture constraint on x + y in factorization!

The Joy of Meetings

- Two people set up a meeting for 12pm
  - Each arrives independently at time uniformly distributed between 12pm and 12:30pm
  - X = # min. past 12pm person 1 arrives   X ~ Uni(0, 30)
  - Y = # min. past 12pm person 2 arrives   Y ~ Uni(0, 30)
  - What is P(first to arrive waits > 10 min. for other)?

$$P(X+10<Y)+P(Y+10<X)=2P(X+10<Y) \text{ by symmetry}$$

$$2P(X+10<Y)=2\iint_{x+10<y} f(x,y)\,dxdy = 2\iint_{x+10<y} f_X(x) f_Y(y)\,dxdy$$

$$=2\int_{y=10}^{30}\int_{x=0}^{y-10}\left(\frac{1}{30}\right)^2 dxdy=\frac{2}{30^2}\int_{y=10}^{30}\left(\int_{x=0}^{y-10}dx\right)dy=\frac{2}{30^2}\int_{y=10}^{30}\left(x\Big|_0^{y-10}\right)dy=\frac{2}{30^2}\int_{y=10}^{30}(y-10)\,dy$$

$$=\frac{2}{30^2}\left(\frac{y^2}{2}-10y\right)\Bigg|_{10}^{30}=\frac{2}{30^2}\left[\left(\frac{30^2}{2}-300\right)-\left(\frac{10^2}{2}-100\right)\right]=\frac{4}{9}$$

Independence of Multiple Variables

- *n* random variables $X_1, X_2, \ldots, X_n$ are called **independent** if:

$$P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n)=\prod_{i=1}^{n} P(X_i=x_i) \quad \text{for all } x_1, x_2, \ldots, x_n$$

- Analogously, for continuous random variables:

$$P(X_1\le a_1, X_2\le a_2, \ldots, X_n\le a_n)=\prod_{i=1}^{n} P(X_i\le a_i) \quad \text{for all } a_1, a_2, \ldots, a_n$$

Independence is Symmetric

- If random variables X and Y independent, then
  - X independent of Y, and
  - Y independent of X
- Duh!? Duh, indeed...
  - Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (I.I.D.) random variables
  - Say $X_n > X_i$ for all $i = 1, \ldots, n - 1$ (i.e. $X_n = \max(X_1, \ldots, X_n)$)
    - Call X_n a "record value" (e.g., record temp. for particular day)
  - Let event A_i = X_i is "record value"
    - Is A_{n+1} independent of A_n?
    - Is A_n independent of A_{n+1}?
    - Easier to answer: Yes!
    - By symmetry, $P(A_i) = 1/n$ for $1 \le i \le n$

Choosing a Random Subset

- From set of *n* elements, choose a subset of size *k* such that all $\binom{n}{k}$ possibilities are <u>equally</u> likely
  - Only have `random()`, which simulates X ~ Uni(0, 1)
- Brute force:
  - Generate all subsets of size *k*
  - Randomly pick one (divide (0, 1) into $\binom{n}{k}$ intervals)
  - Expensive with regard to time and space
  - Bad times!

(Happily) Choosing a Random Subset

- Good times:

```
int indicator(double p) {
   if (random() < p) return 1; else return 0;
}

subset rSubset(k, set of size n) {
   subset_size = 0;
   I[1] = indicator((double)k/n);
   for(i = 1; i < n; i++) {
      subset_size += I[i];
      I[i+1] = indicator((k - subset_size)/(n - i));
   }
   return (subset containing element[i] iff I[i] == 1);
}
```

$P(I[1]=1)=\frac{k}{n}$ and $P(I[i+1]=1 \mid I[1],\ldots,I[i])=\frac{k-\sum_{j=1}^{i} I[j]}{n-i}$ where $1<i<n$

Random Subsets the Happy Way

- Proof (Induction on (k + n)): (i.e., why this algorithm works)
  - Base Case: k = 1, n = 1, Set S = {*a*}, `rSubset` returns {*a*} with $p = 1/\binom{1}{1}$
  - Inductive Hypoth. (IH): for $k + x \le c$, Given set S, |S| = x and $k \le x$, `rSubset` returns any subset S' of S, where |S'| = k, with $p = 1/\binom{x}{k}$
  - Case 1: (where $k + n \le c + 1$) |S| = n (= x + 1), `I[1] = 1`
    - Elem 1 in subset, choose k – 1 elems from remaining n – 1
    - By IH: `rSubset` returns subset S' of size k – 1 with $p = 1/\binom{n-1}{k-1}$
    - P(I[1] = 1, subset S') = $\frac{k}{n}\cdot 1/\binom{n-1}{k-1} = 1/\binom{n}{k}$
  - Case 2: (where $k + n \le c + 1$) |S| = n (= x + 1), `I[1] = 0`
    - Elem 1 not in subset, choose k elems from remaining n – 1
    - By IH: `rSubset` returns subset S' of size k with $p = 1/\binom{n-1}{k}$
    - P(I[1] = 0, subset S') = $\left(1-\frac{k}{n}\right)\cdot 1/\binom{n-1}{k} = \left(\frac{n-k}{n}\right)\cdot 1/\binom{n-1}{k} = 1/\binom{n}{k}$