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)!

**Web Server Requests**

- Let $N = \#$ of requests to web server/day
  - Suppose $N \sim \text{Poi}(\lambda)$
  - Each request comes from a human (probability = $p$) or from a “bot” (probability = $(1 - p)$), independently
  - $X = \#$ requests from humans/day \quad $(X | N) \sim \text{Bin}(N, p)$
  - $Y = \#$ requests from bots/day \quad $(Y | N) \sim \text{Bin}(N, 1 - p)$
  - \(P(X = i, Y = j) = P(X = i, Y = j) \mid X + Y = i + j)P(X + Y = i + j)\)
- Note:
  - $P(X = i, Y = j) = P(X = i, Y = j) \mid X + Y = i + j)P(X + Y = i + j)\)
  - $P(X = i, Y = j) = \binom{i+j}{i} p^i (1-p)^j$
  - $P(X + Y = i + j) = e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$
  - $P(X = i, Y = j) = \binom{i+j}{i} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$

**Independent Continuous Variables**

- Two continuous random variables $X$ and $Y$ are called independent if:
  $$P(X \leq a, Y \leq b) = P(X \leq a) P(Y \leq b) \quad \text{for any } a, b$$
- Equivalently:
  $$f_{X,Y}(a,b) = f_X(a)f_Y(b) \quad \text{for all } a, b$$
- More generally, joint density factors separately:
  $$f_{X,Y}(x,y) = h(x)g(y) \quad \text{where } -\infty < x, y < \infty$$

**Coin Flips**

- Flip coin with probability $p$ of “heads”
  - Flip coin a total of $n + m$ times
  - Let $X = \#$ of heads in first $n$ flips
  - Let $Y = \#$ 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 $Z = \#$ of total heads in $n + m$ flips
  - Are $X$ and $Z$ independent?
    - What if you are told $Z = 0$?

**Web Server Requests (cont.)**

- Let $N = \#$ of requests to web server/day
  - Suppose $N \sim \text{Poi}(\lambda)$
  - Each request comes from a human (probability = $p$) or from a “bot” (probability = $(1 - p)$), independently
  - $X = \#$ requests from humans/day \quad $(X | N) \sim \text{Bin}(N, p)$
  - $Y = \#$ requests from bots/day \quad $(Y | N) \sim \text{Bin}(N, 1 - p)$
  - $P(X = i, Y = j) = e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!} (\lambda(1-p))^j$
    - $= e^{-\lambda} \frac{(dp)^i}{i!} e^{-\lambda} \frac{(d(1-p))^j}{j!} = P(X = i)P(Y = j)$
  - where $X \sim \text{Poi}(\lambda p)$ and $Y \sim \text{Poi}(\lambda(1 - p))$
  - $X$ and $Y$ are independent!

**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 \( X \sim U(0, 30) \)
  - Y = # min. past 12pm person 2 arrives \( Y \sim U(0, 30) \)
  - What is \( P(\text{first to arrive waits > 10 min. for other}) \)?

\[
P(X + 10 < Y) = 2 \int_{x=0}^{10} f_X(x) f_Y(y) \, dy = 2 \int_{y=10}^{30} f_Y(y) \, dy = 2 \frac{y}{30} \bigg|_{y=10}^{30} = \frac{4}{9}
\]

Independence of Multiple Variables

- \( n \) random variables \( X_1, X_2, \ldots, X_n \), are called \( \text{independent if:} \)
  \[
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \quad \text{for all } x_1, x_2, \ldots, x_n
\]
  - Analogously, for continuous random variables:
    \[
P(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n) = \prod_{i=1}^{n} P(X_i \leq a_i) \quad \text{for all } a_1, a_2, \ldots, a_n
\]

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.) continuous random vars
  - 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 \( A_n \) be \( X_n \) is "record value"
    - Is \( A_n \) independent of \( A_m \),? 
    - Easier to answer: Yes!
    - By symmetry, \( P(A_n) = 1/n \)

(Happily) Choosing a Random Subset

- Good times:

```c
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++)
        if (indicator((double)k/(n - subset_size)))
            subset_size += I[1];
    return (subset containing element[i] iff I[1] == 1);
}
PARAM k / 10, n / 20,
P[I[1] = 1] = \frac{k}{n} \quad \text{and} \quad P[I[i+1] = 1 | I[1], I[2], \ldots, I[i]] = \frac{k - \sum_{j=1}^{i} I[j]}{n - \sum_{j=1}^{i} I[j]} \quad \text{where } 1 < i < n
```

Choosing a Random Subset

- From set of \( n \) elements, choose a subset of size \( k \) such that all \( \binom{n}{k} \) possibilities are \( \text{equally likely} \)
  - Only have \( \text{random()} \), which simulates \( X \sim U(0, 1) \)
- Brute force:
  - Generate (an ordering of) 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!

Random Subsets the Happy Way

- Proof (Induction on \( k+n \)): (i.e., why this algorithm works)
  - Base Case: \( k = 1, n = 1 \), \( \text{rSubset returns } \{a\} \) with \( p = \frac{1}{n} \)
  - Inductive Hypoth. (IH): for \( k + x < c \), Given set \( S, |S| = x \) and \( k \leq x \)
    - \( \text{rSubset returns any subset } S' \text{ of } S \text{ where } |S'| = k \text{, with } p = \frac{1}{n} \)
  - Case 1: when \( k + n < c + 1 \)
    - \( |S'| = n - x + 1 \), \( I[1] = 1 \)
      - Elem 1 in subset, choose \( k-1 \) elem from remaining \( n-1 \)
      - By IH: \( \text{rSubset returns subset } S' \text{ of size } k-1 \text{ with } p = \frac{1}{n-1} \)
    - \( P[I[1] = 1, \text{subset } S'] = \frac{k}{n} \frac{1}{n-1} \frac{1}{k-1} \frac{1}{n-1} \)
  - Case 2: when \( k + n < c + 1 \)
    - \( |S'| = n - x + 1 \), \( I[1] = 0 \)
      - Elem 1 not in subset, choose \( k \) elem from remaining \( n-1 \)
      - By IH: \( \text{rSubset returns subset } S' \text{ of size } k \text{ with } p = \frac{1}{n} \)
    - \( P[I[1] = 0, \text{subset } S'] = \left(1 - \frac{1}{n}\right) \frac{n-1}{n} \left(\frac{1}{k-1}\right) (n-1-n) \frac{1}{n-1} \frac{1}{n} \)
### Sum of Independent Binomial RVs

- Let $X$ and $Y$ be independent random variables
  - $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$
  - $X + Y \sim \text{Bin}(n_1 + n_2, p)$

**Intuition:**
- $X$ has $n_1$ trials and $Y$ has $n_2$ trials
- Each trial has same “success” probability $p$
- Define $Z$ to be $n_1 + n_2$ trials, each with success prob. $p$
- $Z \sim \text{Bin}(n_1 + n_2, p)$, and also $Z = X + Y$

- More generally: $X_i \sim \text{Bin}(n_i, p)$ for $1 \leq i \leq N$

\[
\sum_{i=1}^{N} X_i \sim \text{Bin} \left( \sum_{i=1}^{N} n_i, p \right)
\]

### Sum of Independent Poisson RVs

- Let $X$ and $Y$ be independent random variables
  - $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$
  - $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$

**Proof:** (just for reference)
- Rewrite $(X + Y = n)$ as $(X = k, Y = n-k)$ where $0 \leq k \leq n$
- Noting Binomial theorem:
- $P(X + Y = n) = \sum_{k=0}^{n} P(X = k, Y = n-k) = \sum_{k=0}^{n} P(X = k)P(Y = n-k)$
- \[= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} = e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \]
- Noting Binomial theorem:
- $(\lambda_1 + \lambda_2)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$
- \[P(X + Y = n) = e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \]
- $P(X + Y = n) = e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$
- So, $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$