Recall with probability $p$ since requests independent:

Let $X = \#$ of events that occur: $X = \sum_{i=1}^{n} I_i$, $E[X] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} P(A_i)$

Now consider pair of events $A_i, A_j$ occurring

- $I_i I_j = 1$ if both events $A_i$ and $A_j$ occur, 0 otherwise
- Number of pairs of events that occur is $X = \sum_{i \neq j} I_i I_j$

### Let’s Try It with the Binomial

- $X \sim \text{Bin}(n, p)$ $E[X] = \sum_{i=1}^{n} P(A_i) = np$

- Each trial: $X_i \sim \text{Ber}(p)$ $E[X_i] = p$

- Let event $A_i = \text{trial i is success (i.e., } X_i = 1)$

  $E\left[\left(\frac{X}{2}\right)^2\right] = \sum_{i=1}^{n} E[X_i] \cdot \sum_{i=1}^{n} P(A_i, A_i) = \sum_{i=1}^{n} p^2 = \binom{n}{2} p^2$

  $E[X(X - 1)] = E[X^2] - E[X] = n(n-1) p^2$

  $\text{Var}(X) = E[X^2] - (E[X])^2 = (E[X^2]) - E[X] + (E[X])^2$

  $= n(n-1) p^2 + np - (np)^2 = n^2 p^2 - np^2 + np - n^2 p^2$

  $= np(1 - p)$

### Computer Cluster Utilization (cont.)

- Computer cluster with $N$ servers
  - Requests independently go to server $i$ with probability $p_i$
  - Let event $A_i = \text{server i receives no requests}$
  - $X = \#$ of events $A_1, A_2, \ldots, A_N$ that occur
  - $Y = \#$ servers that receive $\geq 1$ request $\Rightarrow N - X$
  - $E[Y]$ after first $n$ requests?
  - Since requests independent: $P(A_i) = (1 - p_i)^r$

  $E[X] = \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} (1 - p_i)^r$

  $E[Y] = N - E[X] = N - \sum_{i=1}^{n} (1 - p_i)^r$

  when $p_i = \frac{1}{N}$ for $1 \leq i \leq N$, $E[Y] = N - \sum_{i=1}^{n} (1 - \frac{1}{N})^r = N [1 - (1 - \frac{1}{N})^r]$

### From Event Pairs to Variance

- Expected number of pairs of events:

  $E\left[\left(\frac{X}{2}\right)^2\right] = \frac{1}{2} E[X^2] - E[I_i] = \frac{1}{2} \sum_{i=1}^{n} P(A_i, A_i)$

  $E[X^2] = 2 \sum_{i=1}^{n} P(A_i, A_i) \Rightarrow E[X^2] = 2 \sum_{i=1}^{n} P(A_i, A_i) + E[X]$  

- Recall: $\text{Var}(X) = E[X^2] - (E[X])^2$

  $\text{Var}(X) = 2 \sum_{i=1}^{n} P(A_i, A_i) + (E[X])^2 - (E[X])^2$

  $= 2 \sum_{i=1}^{n} P(A_i, A_i) + (E[X])^2$ - $\left(\sum_{i=1}^{n} P(A_i)\right)^2$

### Computer Cluster = Coupon Collecting

- Computer cluster with $N$ servers
  - Requests independently go to server $i$ with probability $p_i$
  - Let event $A_i = \text{server i receives no requests}$
  - $X = \#$ of events $A_1, A_2, \ldots, A_N$ that occur
  - $Y = \#$ servers that receive $\geq 1$ request $\Rightarrow N - X$
  - This is really another "Coupon Collector" problem
  - Each server is a "coupon type"
  - Request to server = collecting a coupon of that type

- Hash table version
  - Each server is a bucket in table
  - Request to server = string gets hashed to that bucket
Product of Expectations

- Let X and Y are independent random variables, and g(*) and h(*) are real-valued functions
  \[ E[g(X)]E[h(Y)] = E[g(X)h(Y)] \]

- Proof:
  \[ E[g(X)h(Y)] = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y) \, dx \, dy \]
  \[ = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x)f_X(x)h(y)f_Y(y) \, dx \, dy \]
  \[ = \int_{y=-\infty}^{\infty} h(y)f_Y(y) \, dy \int_{x=-\infty}^{\infty} g(x)f_X(x) \, dx \]
  \[ = E[g(X)]E[h(Y)] \]

The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:
  \[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \]

- Equivalently:
  \[ \text{Cov}(X, Y) = E[XY] - E[X]E[Y] \]

- X and Y independent, E[XY] = E[X]E[Y] → Cov(X, Y) = 0
- But Cov(X, Y) = 0 does not imply X and Y independent!

Another Example of Covariance

Consider the following data:

<table>
<thead>
<tr>
<th>Weight</th>
<th>Height</th>
<th>Weight * Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>57</td>
<td>3708</td>
</tr>
<tr>
<td>71</td>
<td>59</td>
<td>4189</td>
</tr>
<tr>
<td>53</td>
<td>49</td>
<td>2597</td>
</tr>
<tr>
<td>67</td>
<td>62</td>
<td>4154</td>
</tr>
<tr>
<td>55</td>
<td>51</td>
<td>2805</td>
</tr>
<tr>
<td>58</td>
<td>50</td>
<td>2900</td>
</tr>
<tr>
<td>77</td>
<td>55</td>
<td>4235</td>
</tr>
<tr>
<td>57</td>
<td>48</td>
<td>2736</td>
</tr>
<tr>
<td>56</td>
<td>42</td>
<td>2352</td>
</tr>
<tr>
<td>51</td>
<td>42</td>
<td>2142</td>
</tr>
<tr>
<td>61</td>
<td>61</td>
<td>3776</td>
</tr>
</tbody>
</table>

\[ = 3355.83 - (62.75)(52.75) \]
\[ = 45.77 \]

Properties of Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:
  \[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \]

- Equivalently:
  \[ \text{Cov}(X, Y) = E[XY] - E[X]E[Y] \]

- Covariance of sums of random variables
  \[ \text{Cov}(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} \text{Cov}(X_i, Y_j) \]
Variance of Sum of Variables

- \[ \text{Var} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{Var} (X_i) + 2 \sum_{j=1}^{n} \sum_{i<j} \text{Cov} (X_i, X_j) \]

- **Proof:**
  \[ \text{Var} \left( \sum_{i=1}^{n} X_i \right) = \text{Cov} \left( \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i \right) \]
  \[ = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} (X_i, X_j) \]
  \[ = \sum_{i=1}^{n} \text{Var} (X_i) + \sum_{i<j} \sum_{j<i} \text{Cov} (X_i, X_j) \]
  \[ \text{By symmetry:} \quad \text{Cov} (X_i, X_j) = \text{Cov} (X_j, X_i) \]
  \[ = \sum_{i=1}^{n} \text{Var} (X_i) + 2 \sum_{j=1}^{n} \sum_{i<j} \text{Cov} (X_i, X_j) \]
  \[ = \sum_{i=1}^{n} \text{Var} (X_i) + 2 \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} \text{Cov} (X_i, X_j) \]
  \[ \text{If all } X_i \text{ and } X_j \text{ independent } (i \neq j): \text{Var} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{Var} (X_i) \]

- **Note:** \( \text{Cov} (X, X) = \text{Var} (X) \)

***Holá Compadre: La Distribución Binomial***

- Let \( Y \sim \text{Bin}(n, p) \)
  - \( n \) independent trials
  - Let \( X_i = 1 \) if \( i \)-th trial is “success”, 0 otherwise
  - \( X_i \sim \text{Ber}(p) \)
  - \( \text{E}[X] = p \)
  - \( \text{Var}(Y) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) \)
  - \( \text{Var}(X) = \text{E}[X^2] - (\text{E}[X])^2 \) since \( X_i^2 = X_i \)
  - \( p - p^2 = p(1-p) \)
  - \( \text{Var}(Y) = n \text{Var}(X) = np(1-p) \)

***Variance of Sample Mean***

- Consider \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \)
  - \( X_i \) have distribution \( F \) with \( \text{E}[X_i] = \mu \) and \( \text{Var}(X_i) = \sigma^2 \)
  - We call sequence of \( X_i \) a **sample** from distribution \( F \)
  - Recall sample mean: \( \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \) where \( \text{E} [\bar{X}] = \mu \)
  - What is \( \text{Var}(\bar{X}) \)?
  \[ \text{Var}(\bar{X}) = \text{Var} \left( \frac{\sum_{i=1}^{n} X_i}{n} \right) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^{n} X_i \right) \]
  \[ = \frac{1}{n^2} \left( \frac{1}{n} \right)^2 \sum_{i=1}^{n} \text{Var} (X_i) \]
  \[ = \frac{1}{n} \sigma^2 \]

***Sample Variance***

- Consider \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \)
  - \( X_i \) have distribution \( F \) with \( \text{E}[X_i] = \mu \) and \( \text{Var}(X_i) = \sigma^2 \)
  - We call sequence of \( X_i \) a **sample** from distribution \( F \)
  - Recall sample mean: \( \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \) where \( \text{E} [\bar{X}] = \mu \)
  - Sample deviation: \( X_i - \bar{X} \), for \( i = 1, 2, \ldots, n \)
  - Sample variance: \( s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \)
  - What is \( \text{E}[s^2] \)?
  - \( \text{E}[s^2] = \sigma^2 \)
  - We say \( s^2 \) is “unbiased estimate” of \( \sigma^2 \)

***Proof that \( \text{E}[S^2] = \sigma^2 \) (just for reference)***

\[ \text{E}[S^2] = \text{E} \left[ \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \right] = (n-1) \text{E}[S^2] = \text{E} \left[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] \]

\[ (n-1) \text{E}[S^2] = \text{E} \left[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = \text{E} \left[ \sum_{i=1}^{n} (X_i - \mu)^2 + n(\bar{X} - \mu)^2 \right] \]

\[ = \text{E} \left[ \sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i=1}^{n} (\bar{X} - \mu)^2 + 2 \sum_{i=1}^{n} (X_i - \mu)(\bar{X} - \mu) \right] \]

\[ = \text{E} \left[ \sum_{i=1}^{n} (X_i - \mu)^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) \right] \]

\[ = \text{E} \left[ \sum_{i=1}^{n} (X_i - \mu)^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu) n(\bar{X} - \mu) \right] \]

\[ = n\sigma^2 + n \text{Var}(\bar{X}) = n\sigma^2 + n \frac{\sigma^2}{n} - n\sigma^2 + \sigma^2 = (n-1)\sigma^2 \]

\[ \text{So,} \quad \text{E}[S^2] = \sigma^2 \]