13: Statistics on Multiple Random Variables

Jerry Cain
February 7, 2024

Lecture Discussion on Ed
Coupon Collecting
Coupon collecting and server requests

The **coupon collector’s problem** in probability theory:
- You buy boxes of cereal.
- There are $k$ different types of coupons
- For each box you buy, you "collect" a coupon of type $i$.

1. How many coupons do you expect after buying $n$ boxes of cereal?

What is the expected number of servers utilized after $n$ requests?

* 52% of Amazon profits
** more profitable than Amazon’s North America commerce operations

source
Computer cluster utilization

Consider a computer cluster with \( k \) servers. We send \( n \) requests.

- Requests independently go to server \( i \) with probability \( p_i \)
- Let \( X = \# \) servers that receive \( \geq 1 \) request.

What is \( E[X] \)?
Computer cluster utilization

Consider a computer cluster with \( k \) servers. We send \( n \) requests.

- Requests independently go to server \( i \) with probability \( p_i \)
- Let \( X = \# \) servers that receive \( \geq 1 \) request.

What is \( E[X] \)?

1. Define additional random variables.

Let:
- \( A_i = \) event that server \( i \) receives \( \geq 1 \) request
- \( X_i = \) indicator for \( A_i \)
  - \( X_i = 1 \) if \( A_i \) holds
  - \( X_i = 0 \) if \( A_i \) doesn't hold

\[
P(A_i) = 1 - P(\text{no requests to } i) = 1 - (1 - p_i)^n
\]

\( E[X_i] = P(A_i) = 1 - (1 - p_i)^n \)

2. Solve.

\[
E[X] = E \left[ \sum_{i=1}^{k} X_i \right] = \sum_{i=1}^{k} E[X_i] = \sum_{i=1}^{k} (1 - (1 - p_i)^n)
\]

Note: \( A_i \) are dependent!
Coupon collecting problems: Hash tables

The **coupon collector’s problem** in probability theory:

- You buy boxes of cereal.
- There are $k$ different types of coupons
- For each box you buy, you ”collect” a coupon of type $i$.

1. How many coupons do you expect after buying $n$ boxes of cereal?

2. How many boxes do you expect to buy until you have one of each coupon?

What is the expected number of utilized servers after $n$ requests?

What is the expected number of strings to hash until each bucket has $\geq 1$ string?
Hash Tables

Consider a hash table with \( k \) buckets.

- Strings are equally likely to get hashed into any bucket (independently).
- Let \( Y = \# \) strings to hash until each bucket \( \geq 1 \) string.

What is \( E[Y] \)?

1. Define additional random variables. How should we define \( Y_i \) such that \( Y = \sum_i Y_i \)?

2. Solve.
Hash Tables

Consider a hash table with $k$ buckets.

- Strings are equally likely to get hashed into any bucket (independently).
- Let $Y$ = # strings to hash until each bucket $\geq 1$ string.

What is $E[Y]$?

1. Define additional random variables.

Let: $Y_i$ = # of trials needed to get success after $i$-th success
- Success: hash string to previously empty bucket
- If $i$ non-empty buckets: $P(\text{success}) = \frac{k-i}{k}$

\[ P(Y_i = n) = \left( \frac{i}{k} \right)^{n-1} \left( \frac{k-i}{k} \right) \]

Equivalently, $Y_i \sim \text{Geo} \left( p = \frac{k-i}{k} \right)$

\[ E[Y_i] = \frac{1}{p} = \frac{k}{k-i} \]
Hash Tables

Consider a hash table with $k$ buckets.

- Strings are equally likely to get hashed into any bucket (independently).
- Let $Y = \#$ strings to hash until each bucket $\geq 1$ string.

What is $E[Y]$?

1. Define additional random variables. Let: $Y_i = \#$ of trials to needed get success after $i$-th success

   $$Y_i \sim \text{Geo}\left(p = \frac{k - i}{k}\right), \quad E[Y_i] = \frac{1}{p} = \frac{k}{k - i}$$

2. Solve. $Y = Y_0 + Y_1 + \cdots + Y_{k-1}$

   $$E[Y] = E[Y_0] + E[Y_1] + \cdots + E[Y_{k-1}]$$

   $$= \frac{k}{k} + \frac{k}{k - 1} + \frac{k}{k - 2} + \cdots + \frac{k}{1} = k \left[ \frac{1}{k} + \frac{1}{k - 1} + \cdots + 1 \right] = O(k \log k)$$
Covariance
Statistics of sums of RVs

For any random variables $X$ and $Y$,

$$E[X + Y] = E[X] + E[Y]$$

Var($X + Y$) = ?

But first, a new statistic!
Spot the difference

Compare/contrast the following two distributions:

Both distributions have the same $E[X]$, $E[Y]$, $\text{Var}(X)$, and $\text{Var}(Y)$

Difference: how the two variables vary with each other.
Covariance

The covariance of two variables $X$ and $Y$ is:

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

Proof of second part (rewriting $E[X]$, $E[Y]$ as $\mu_X$, $\mu_Y$ to emphasize the fact they’re each constants):

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[(X - \mu_X)(Y - \mu_Y)]
= E[XY] - \mu_Y X - \mu_X Y + \mu_X \mu_Y
= E[XY] - E[\mu_Y X] - E[\mu_X Y] + E[\mu_X \mu_Y]
= E[XY] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y
= E[XY] - \mu_X \mu_Y = E[XY] - E[X]E[Y]$$

(linearity of expectation)
($\mu_X$, $\mu_Y$ are constants)
Covariance

The covariance of two variables $X$ and $Y$ is:

\[
\]

Covariance measures how one random variable varies with a second.
- Outside temperature and utility bills have a negative covariance.
- Handedness and musical ability have near zero covariance.
- Product demand and price have a positive covariance.
Feel the covariance

Is the covariance positive, negative, or zero?

1. 
   \[ Y = y \]
   \[ X = x \]
   \[ E[X] \]
   \[ E[Y] \]

2. 
   \[ Y = y \]
   \[ X = x \]
   \[ E[X] \]
   \[ E[Y] \]

3. 
   \[ Y = y \]
   \[ X = x \]
   \[ E[X] \]
   \[ E[Y] \]

\[
\]
Feel the covariance

Is the covariance positive, negative, or zero?

1. Positive
   - as $x$ increases, so does $y$; positive covariance

2. Negative
   - as $x$ goes up, $y$ goes down; negative covariance

3. Zero
   - no obvious pattern in how $y$ changes as $x$ increases;
   - no covariance

\[
\]
Covarying humans

<table>
<thead>
<tr>
<th>Weight (kg)</th>
<th>Height (in)</th>
<th>W \cdot H</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>57</td>
<td>3648</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>76</td>
<td>61</td>
<td>4636</td>
</tr>
<tr>
<td>68</td>
<td>57</td>
<td>3876</td>
</tr>
</tbody>
</table>

$E[W] = 62.75$

$E[H] = 52.75$

$E[WH] = 3355.83$

What is the covariance of weight $W$ and height $H$?


$$= 3355.83 - (62.75)(52.75)$$

(positive) $= 45.77$

Covariance $> 0$: one variable ↑, other variable ↑
Properties of Covariance

The covariance of two variables $X$ and $Y$ is:

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

Properties:

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2. $\text{Var}(X) = E[X^2] - (E[X])^2 = E[XX] - E[X]E[X] = \text{Cov}(X, X)$
3. Covariance of sums = sum of all pairwise covariances
   $\text{Cov}(X_1 + X_2, Y_1 + Y_2) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_2, Y_1) + \text{Cov}(X_1, Y_2) + \text{Cov}(X_2, Y_2)$
4. Covariance under linear transformation: $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$
   (proof left to you)
Zero covariance does not imply independence

Let $X$ take on values $\{-1, 0, 1\}$ with equal probability $1/3$.

Define $Y = \begin{cases} 
1 & \text{if } X = 0 \\
0 & \text{otherwise}
\end{cases}$

What is the joint PMF of $X$ and $Y$?
Zero covariance does not imply independence

Let $X$ take on values $\{-1, 0, 1\}$ with equal probability $1/3$.

Define $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

1. $E[X] = \quad E[Y] =$

2. $E[XY] =$

3. $\text{Cov}(X, Y) =$

4. Are $X$ and $Y$ independent?
### Zero covariance does not imply independence

Let $X$ take on values $\{-1, 0, 1\}$ with equal probability $1/3$. Define $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$-1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>$1$</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

**Marginal PMF of $X$, $p_X(x)$**

<table>
<thead>
<tr>
<th>$Y$</th>
<th>0</th>
<th>1/3</th>
<th>1/3</th>
<th>1/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
<td>2/3</td>
</tr>
<tr>
<td>$1$</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
</tr>
</tbody>
</table>

**Marginal PMF of $Y$, $p_Y(y)$**

1. $E[X] = -1 \left(\frac{1}{3}\right) + 0 \left(\frac{1}{3}\right) + 1 \left(\frac{1}{3}\right) = 0 \quad E[Y] = 0 \left(\frac{2}{3}\right) + 1 \left(\frac{1}{3}\right) = 1/3$

2. $E[XY] = (-1 \cdot 0) \left(\frac{1}{3}\right) + (0 \cdot 1) \left(\frac{1}{3}\right) + (1 \cdot 0) \left(\frac{1}{3}\right)$
   
   $= 0$

   
   $= 0 - 0(1/3) = 0$

4. **Are $X$ and $Y$ independent?**

   $P(Y = 0|X = 1) = 1$
   
   $\neq P(Y = 0) = 2/3$
Variance of sums of RVs
Statistics of sums of RVs

For any random variables $X$ and $Y$,

$$E[X + Y] = E[X] + E[Y]$$

$$Var(X + Y) = Var(X) + 2 \cdot Cov(X, Y) + Var(Y)$$
**Variance of general sum of RVs**

For any random variables $X$ and $Y$,

\[
\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)
\]

Proof:

\[
\text{Var}(X + Y) = \text{Cov}(X + Y, X + Y) = \text{Cov}(X, X) + \text{Cov}(X, Y) + \text{Cov}(Y, X) + \text{Cov}(Y, Y)
\]

\[
= \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)
\]

More generally:

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

(proof in extra slides)
Statistics of sums of RVs

For any random variables $X$ and $Y$,

$$E[X + Y] = E[X] + E[Y]$$

$$\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$$

For **independent** $X$ and $Y$,

$$E[XY] = E[X]E[Y]$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

(Lemma: proof in extra slides)
Variance of sum of independent RVs

For independent \( X \) and \( Y \),

\[
\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)
\]

Proof:

   \[= 0 \]

   \( X \) and \( Y \) are independent

2. \( \text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y) \)
   \[= \text{Var}(X) + \text{Var}(Y) \]

\( X \) and \( Y \) are independent

**NOT bidirectional:**
\( \text{Cov}(X, Y) = 0 \) does NOT imply independence of \( X \) and \( Y \)!
Proving Variance of the Binomial

\[ X \sim \text{Bin}(n, p) \quad \text{Var}(X) = np(1 - p) \]

To simplify the algebra a bit, let \( q = 1 - p \), so \( p + q = 1 \).

So:

\[
\begin{align*}
E(X^2) &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} \\
&= np + np^2 + npq \\
&= np + np^2 + np(1 - p) \\
&= np + np^2 - npq \\
&= np + npq \\
&= np + np(1 - q) \\
&= np + np(1 - (1 - p)) \\
&= np + np \\
&= np + np(1 - p) \\
&= np + npq \\
&= np(1 + q) \\
&= np.
\end{align*}
\]

Definition of Binomial Distribution: \( p + q = 1 \)

Factors of Binomial Coefficient: \( \binom{n}{k} = \binom{n}{k-1} \)

Change of limit; term is zero when \( k = 0 \)

Putting \( j = k - 1, m = n - 1 \)

Splitting sum up into two

Factors of Binomial Coefficient: \( \binom{n}{m} = \binom{n}{m-1} \)

Change of limit; term is zero when \( j = n \)

Binomial Theorem

so \( p + q = 1 \)

by algebra.

Let’s instead prove this using independence and variance!

proofwiki.org
Proving Variance of the Binomial

\[ X \sim \text{Bin}(n, p) \quad \text{Var}(X) = np(1 - p) \]

Let \[ X = \sum_{i=1}^{n} X_i \]

Let \( X_i \) = \( i \)th trial is heads
\( X_i \sim \text{Ber}(p) \)
\( \text{Var}(X_i) = p(1 - p) \)

\( X_i \) are independent (by definition)

\[ \text{Var}(X) = \text{Var} \left( \sum_{i=1}^{n} X_i \right) \]
\[ = \sum_{i=1}^{n} \text{Var}(X_i) \]
\[ = \sum_{i=1}^{n} p(1 - p) \]
\[ = np(1 - p) \]

\( X_i \) are independent, therefore variance of sum = sum of variance

Variance of Bernoulli
Correlation
Covarying humans

What is the covariance of weight $W$ and height $H$?


$$= 3355.83 - (62.75)(52.75)$$

$$= 45.77 \text{ (positive)}$$

What about weight (lb) and height (cm)?

$$\text{Cov}(2.20W, 2.54H)$$

$$= E[2.20W \cdot 2.54H] - E[2.20W]E[2.54H]$$

$$= 18752.38 - (138.05)(133.99)$$

$$= 255.06 \text{ (positive)}$$

⚠ Covariance depends on units!

Sign of covariance (+/-) more meaningful than magnitude
Correlation

The correlation of two variables $X$ and $Y$ is:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Note: $-1 \leq \rho(X, Y) \leq 1$
- Correlation measures the linear relationship between $X$ and $Y$:
  - $\rho(X, Y) = 1 \implies Y = aX + b$, where $a = \sigma_Y / \sigma_X$
  - $\rho(X, Y) = -1 \implies Y = aX + b$, where $a = -\sigma_Y / \sigma_X$
  - $\rho(X, Y) = 0 \implies$ uncorrelated (absence of linear relationship)
Correlation reps

What is the correlation coefficient $\rho(X,Y)$?

1. $\rho(X,Y) = 1$
2. $\rho(X,Y) = -1$
3. $\rho(X,Y) = 0$
4. Other (by yourself)
Correlation reps

What is the correlation coefficient $\rho(X,Y)$?

1. [Graph showing a negative linear relationship]  
   B. $\rho(X,Y) = -1$  
   $Y = -aX + b$  
   $a > 0$

2. [Graph showing a positive linear relationship]  
   A. $\rho(X,Y) = 1$  
   $Y = aX + b$  
   $a > 0$

3. [Graph showing no linear relationship]  
   C. $\rho(X,Y) = 0$  
   “uncorrelated”

4. [Graph showing a non-linear relationship]  
   C. $\rho(X,Y) = 0$  
   $Y = X^2$

$X$ and $Y$ can be nonlinearly related even if $\rho(X,Y) = 0$. 

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Winter 2024
Throwback to CS103: Conditional statements

Statement $P \rightarrow Q$: Independence $\Rightarrow$ No correlation

Contrapositive $\neg Q \rightarrow \neg P$: Correlation $\Rightarrow$ Dependence

Inverse $\neg P \rightarrow \neg Q$: Dependence $\Rightarrow$ Correlation

Converse $Q \rightarrow P$: No correlation $\Rightarrow$ Independence

"Correlation does not imply causation"
Extras
Expectation of product of independent RVs

If $X$ and $Y$ are independent, then

$$E[XY] = E[X]E[Y]$$
$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof:

$$E[g(X)h(Y)] = \sum_y \sum_x g(x)h(y)p_{X,Y}(x, y)$$

$$= \sum_y \sum_x g(x)h(y)p_X(x)p_Y(y)$$

$$= \sum_y \left( h(y)p_Y(y) \sum_x g(x)p_X(x) \right)$$

$$= \left( \sum_x g(x)p_X(x) \right) \left( \sum_y h(y)p_Y(y) \right)$$

$$= E[g(X)]E[h(Y)]$$

(for continuous proof, replace summations with integrals)

$X$ and $Y$ are independent

Terms dependent on $y$ are constant in integral of $x$

Summations separate
**Variance of Sums of Variables**

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

**Proof:**

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

\[
= \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \text{Cov}(X_i, X_j)
\]

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

- Symmetry of covariance: \(\text{Cov}(X, X) = \text{Var}(X)\)
- Adjust summation bounds