13: Statistics of Multiple RVs

Lisa Yan and Jerry Cain
October 12, 2020
### Quick slide reference

<table>
<thead>
<tr>
<th>Page</th>
<th>Topic</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Expectation of Common RVs</td>
<td>13a_expectation_sum</td>
</tr>
<tr>
<td>8</td>
<td>Coupon Collecting Problems</td>
<td>13b_coupon_collecting</td>
</tr>
<tr>
<td>14</td>
<td>Covariance</td>
<td>13c_covariance</td>
</tr>
<tr>
<td>20</td>
<td>Independence and Variance</td>
<td>13d_variance_sum</td>
</tr>
<tr>
<td>27</td>
<td>Exercises</td>
<td>LIVE</td>
</tr>
<tr>
<td>48</td>
<td>Correlation</td>
<td>LIVE</td>
</tr>
</tbody>
</table>
Expectation of Common RVs
Linearity of Expectation is useful

Expectation is a linear mathematical operation. If $X = \sum_{i=1}^{n} X_i$:

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

• Even if you don’t know the **distribution** of $X$ (e.g., because the joint distribution of $(X_1, \ldots, X_n)$ is unknown), you can still compute **expectation** of $X$!!

• Problem-solving key: Define $X_i$ such that $X = \sum_{i=1}^{n} X_i$

**Most common use cases:**
- $E[X_i]$ easy to calculate
- Or sum of dependent RVs
Expectations of common RVs: Binomial

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

# of successes in \( n \) independent trials with probability of success \( p \)

Recall: \( \text{Bin}(1, p) = \text{Ber}(p) \)

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

Let \( X_i = i^{th} \) trial is heads \( X_i \sim \text{Ber}(p) \), \( E[X_i] = p \)

\[ E[X] = E \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np \]
Expectations of common RVs: Negative Binomial

\[ Y \sim \text{NegBin}(r, p) \quad E[Y] = \frac{r}{p} \]

# of independent trials with probability of success \( p \) until \( r \) successes

Recall: \( \text{NegBin}(1, p) = \text{Geo}(p) \)

\[ Y = \sum_{i=1}^{?} Y_i \]

1. How should we define \( Y_i \)?

2. How many terms are in our summation?
Expectations of common RVs: Negative Binomial

\[ Y \sim \text{NegBin}(r, p) \quad E[Y] = \frac{r}{p} \]

# of independent trials with probability of success \( p \) until \( r \) successes

Recall: \( \text{NegBin}(1, p) = \text{Geo}(p) \)

Let \( Y_i \) = # trials to get \( i \)th success (after \((i-1)\)th success)

\[ Y_i \sim \text{Geo}(p), \quad E[Y_i] = \frac{1}{p} \]

\[
E[Y] = E \left[ \sum_{i=1}^{r} Y_i \right] = \sum_{i=1}^{r} E[Y_i] = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}
\]
Coupon Collecting Problems
Linearity of Expectation is useful

Expectation is a linear mathematical operation. If $X = \sum_{i=1}^{n} X_i$ :

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

• Even if you don’t know the distribution of $X$ (e.g., because the joint distribution of $(X_1, ..., X_n)$ is unknown), you can still compute expectation of the sum!!

• Problem-solving key: Define $X_i$ such that $X = \sum_{i=1}^{n} X_i$

Most common use cases:
• $E[X_i]$ easy to calculate
• Or sum of dependent RVs
Coupon collecting problems: 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 utilized servers after $n$ requests?

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

source

Lisa Yan and Jerry Cain, CS109, 2020
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 \( A_i \)

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

Note: \( A_i \) are dependent!

2. Solve.

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

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

\[
= \sum_{i=1}^{k} 1 - \sum_{i=1}^{k} (1 - p_i)^n = k - \sum_{i=1}^{k} (1 - p_i)^n
\]
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?

Stay tuned for live lecture!
Covariance
Statistics of sums of RVs

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

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

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

But first... a new statistic!
Spot the difference

Compare/contrast the following two distributions:

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

Difference: how the two variables vary with each other.

Assume all points are equally likely.

$$P(X = x, Y = y) = \frac{1}{N}$$
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:

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

(linearity of expectation)

$(E[X], E[Y]$ are scalars)
Covarying humans

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. $\text{Var}(X) = E[X^2] - (E[X])^2 = \text{Cov}(X, X)$
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. Non-linearity (to be discussed in live lecture)
Variance of sums of RVs
Statistics of sums of RVs

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

\[ E[X + Y] = E[X] + E[Y] \]

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

   
   
   $$= 0$$

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

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

Let’s instead prove this using independence and variance!

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

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

So,

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

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

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

Change of limit: term is zero when \( k = 1 \)

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

Splitting sum up into two

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

Change of limit: term is zero when \( j = 0 \)

Binomial Theorem

So \( p + q = 1 \)

By algebra,

Then,

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

Expectation of Binomial Distribution: \( \mathbb{E}(X) = np \)

as required.
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 \( X_i = \text{ith trial is heads} \)
\( X_i \sim \text{Ber}(p) \)
\( \text{Var}(X_i) = p(1 - p) \)

\( X_i \) 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
13: Statistics of Multiple RVs

Lisa Yan and Jerry Cain
October 12, 2020
Where are we now? A roadmap of CS109

Last week: Joint distributions
\[ p_{X,Y}(x,y) \]

Today: Statistics of multiple RVs!
\[ \text{Var}(X + Y) \]
\[ E[X + Y] \]
\[ \text{Cov}(X, Y) \]
\[ \rho(X, Y) \]

Wednesday: Conditional distributions
\[ p_{X|Y}(x|y) \]
\[ E[X|Y] \]

Friday: Modeling with Bayesian Networks

Lisa Yan and Jerry Cain, CS109, 2020
Don’t we already know linearity of expectation?

Expectation is a linear mathematical operation. If \( X = \sum_{i=1}^{n} X_i \) :

\[
E[X] = E \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i]
\]

We covered this back in Lecture 6 (when we first learned expectation)!
- Proved binomial: sum of 1s or 0s
- Hat check (section): sum of 1s or 0s
- We ignored (in)dependence of events.

Why are we learning this again?
- Well, now we can prove it!
- We can now ignore any random variables dependencies!
- Our approach is still the same!
Proof of expectation of a sum of RVs

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

\[ = \sum_x \sum_y xp_{X,Y}(x, y) + \sum_x \sum_y yp_{X,Y}(x, y) \]

\[ = \sum_x x \sum_y p_{X,Y}(x, y) + \sum_y y \sum_x p_{X,Y}(x, y) \]

\[ = \sum_x xp_X(x) + \sum_y yp_Y(y) \]

\[ = E[X] + E[Y] \]

\[ E[X + Y] = E[X] + E[Y] \]

LOTUS,
\[ g(X, Y) = X + Y \]

Linearity of summations (and integrals, btw)

Marginal PMFs for \( X \) and \( Y \)
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?
Check out the properties on the next slide (Slide 33). Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/146231

Breakout rooms: 4 min. Introduce yourself!
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 to get success after $i$-th success
     - Success: hash string to previously empty bucket
     - If $i$ non-empty buckets: $P(\text{success}) = \frac{k - i}{k}$

2. Solve.

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

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

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.
Slide 38 has a question to go over by yourself.

Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/146231

Think by yourself: 1 min
Feel the covariance

Is the covariance positive, negative, or zero?

\[
\]
Feel the covariance

Is the covariance positive, negative, or zero?

1. \( E[X] \)
   \( E[Y] \)
   positive

2. \( E[X] \)
   \( E[Y] \)
   negative

3. \( E[X] \)
   \( E[Y] \)
   zero

Cov\((X, Y)\) = \( E[(X - E[X])(Y - E[Y])] \)
= \( E[XY] - E[X]E[Y] \)
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. $\text{Var}(X) = \text{Cov}(X, X)$
3. $\text{Cov}\left(\sum_i X_i, \sum_j Y_j\right) = \sum_i \sum_j \text{Cov}(X_i, Y_j)$
4. $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y) + b$ \(\times\)

Covariance is non-linear: $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$
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 $Y$,

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

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

\text{(Lemma: proof in extra slides)}

$\text{Cov}(X, Y) = 0$ does NOT imply independence of $X$ and $Y$!
Zero covariance does not imply independence

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

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

What is the joint PMF of $X$ and $Y$?
Check out the properties on the next slide (Slide 44). Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/146231

Breakout rooms: 4 min. Introduce yourself!
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>0/3</td>
<td>0/3</td>
<td>0/3</td>
</tr>
<tr>
<td>$1$</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td>Sum</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>$1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td>$1$</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>Sum</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

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

1. $E[X] = \frac{-1}{3} \cdot 0 + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$
   $E[Y] = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = 1/3$

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

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

4. Are $X$ and $Y$ independent? $\times$

   $P(Y = 0 | X = 1) = 1$
   $\neq P(Y = 0) = 2/3$

Lisa Yan and Jerry Cain, CS109, 2020

Stanford University
Interesting probability news

**Probability and Game Theory in *The Hunger Games***

“Suppose the parents in a given district gave birth to only...five girls, and that all of these kids were born at the same time.”

- Not a probability mass function
- Also duh? (P(you get chosen if you’re the only person) = 1)
- You now know enough Python/ probability to write a better simulation to model the Reaping!!!!
- (game theory part of the article is good)

Topical book review! Fiction is brain food.

Rochester author takes scary look at Big Pharma in debut novel

- "Called 'Malcharist,' it is a completely made-up story about a potentially dangerous drug being put on the market — with outsourced drug trial research, ghostwritten studies, lack of access to raw drug-trial data, and doctors essentially paid to champion new drugs."

- "[Paul John] Scott’s novel is actually a thriller, with not-quite-believable villains who need to be exposed. Yet it’s too wonky to be a beach read. There’s even a conversation over the [😍] probability concept of p-values [😍]."

- "Scott takes his writer into one of those medical meetings he once found so cool, and his book reproduces enough of the numbers — yes, [😊] number tables [😊] in a thriller — that the reader can see the fictional speaker’s good point that the data really do give up their secrets."

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 $a = -\sigma_Y / \sigma_X$
  - $\rho(X, Y) = 0 \implies \text{"uncorrelated" (absence of linear relationship)}$
Slide 52 has a question to go over by yourself.

Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/146231

Think by yourself: 1 min
Correlation reps

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

1. 

2. 

3. 

4. 

A. $\rho(X,Y) = 1$
B. $\rho(X,Y) = -1$
C. $\rho(X,Y) = 0$
D. Other
Correlation reps

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

1. B. $\rho(X, Y) = -1$
   
   $Y = -aX + b$
   
   $a > 0$

2. A. $\rho(X, Y) = 1$
   
   $Y = aX + b$
   
   $a > 0$

3. C. $\rho(X, Y) = 0$
   
   “uncorrelated”

4. C. $\rho(X, Y) = 0$
   
   $Y = X^2$

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

\( \rho(X, Y) \) is used a lot to statistically quantify the relationship b/t X and Y.

Correlation: 0.947091
Spurious Correlations

\( \rho(X, Y) \) is used a lot to statistically quantify the relationship b/t X and Y.

Correlation:

0.947091

Per capita cheese consumption correlates with Number of people who died by becoming tangled in their bedsheets.
Divorce vs. Margarine

Source: US Census, USDA, tylervigen.com


Lisa Yan and Jerry Cain, CS109, 2020

Stanford University
Arcade revenue vs. CS PhDs

Total revenue generated by arcades correlates with Computer science doctorates awarded in the US

Correlation: 0.947091

Data sources: U.S. Census Bureau and National Science Foundation

Lisa Yan and Jerry Cain, CS109, 2020
Extra
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:

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

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

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

Symmetry of covariance: \(\text{Cov}(X, X) = \text{Var}(X)\)

Adjust summation bounds