# Chapter 5. Multiple Random Variables

## 5.4: Covariance and Correlation

Slides (Google Drive) Alex Tsun Video (YouTube)

In this section, we'll learn about covariance; which as you might guess, is related to variance. It is a function of two random variables, and tells us whether they have a positive or negative linear relationship. It also helps us finally compute the variance of a sum of *dependent* random variables, which we have not yet been able to do.

### 5.4.1 Covariance and Properties

We will start with the definition of covariance: $\mathsf{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. By LOTUS, we know this is equal to (where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$)

$$\sum_x \sum_y (x - \mu_X)(y - \mu_Y) p_{X,Y}(x, y)$$

Intuitively, we can see the following possibilities:

- $x > \mu_X, y > \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) > 0$ ($X, Y$ both above their means)
- $x < \mu_X, y < \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) > 0$ ($X, Y$ both below their means)
- $x < \mu_X, y > \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) < 0$ ($X$ below its mean, $Y$ above its mean)
- $x > \mu_X, y < \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) < 0$ ($X$ above its mean, $Y$ below its mean)

So we get a weighted average (by $p_{X,Y}$) of these positive or negative quantities. Just with this brief intuition, we can say that covariance is positive when $X, Y$ are usually both above/below their means, and negative if they are opposite. That is, covariance is positive in general when increasing one variable leads to an increase in the other, and negative when increasing one variable leads to a decrease in the other.

**Definition 5.4.1: Covariance**

Let $X, Y$ be random variables. The **covariance** of $X$ and $Y$ is:

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

This should remind you of the definition of variance - think of replacing $Y$ with $X$ and you'll see it!
Note: Covariance can be negative, unlike variance.

Covariance satisfies the following properties:

1. If $X \perp Y$, then $\mathsf{Cov}(X,Y) = 0$ (but not necessarily vice versa, because the covariance could be zero but $X$ and $Y$ could not be independent).
2. $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$. (Just plug in $Y = X$).

3. $\mathsf{Cov}(X, Y) = \mathsf{Cov}(Y, X)$. (Multiplication is commutative).
4. $\mathsf{Cov}(X + c, Y) = \mathsf{Cov}(X, Y)$. (Shifting doesn't and shouldn't affect the covariance).
5. $\mathsf{Cov}(aX + bY, Z) = a \cdot \mathsf{Cov}(X, Z) + b \cdot \mathsf{Cov}(Y, Z)$. This can be easily remembered like the distributive property of scalars $(aX + bY)Z = a(XZ) + b(YZ)$.
6. $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X, Y)$, and hence if $X \perp Y$, then $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$ (as we discussed earlier).
7. $\mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathsf{Cov}(X_i, Y_j)$. That is covariance works like FOIL (first, outer, inner, last) for multiplication of sums ($(a + b + c)(d + e) = ad + ae + bd + be + cd + ce$).

*Proof of Covariance Alternate Formula.* We will prove that $\mathsf{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

$$\begin{aligned}
\mathsf{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] && \text{[def of covariance]} \\
&= \mathbb{E}[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]] && \text{[algebra]} \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] && \text{[Linearity of Expectation]} \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && \text{[algebra]}
\end{aligned}$$

□

*Proof of Property 1: Covariance of Independent RVs is* 0.

We actually proved in 5.1 already that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ when $X, Y$ are independent. Hence,

$$\mathsf{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

□

*Proof of Property 6: Variance of Sum of RVs.*

We will show that in general, for any RVs $X$ and $Y$, that

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

$$\begin{aligned}
\mathsf{Var}(X + Y) &= \mathsf{Cov}(X + Y, X + Y) && \text{[covariance with self = variance]} \\
&= \mathsf{Cov}(X, X) + \mathsf{Cov}(X, Y) + \mathsf{Cov}(Y, X) + \mathsf{Cov}(Y, Y) && \text{[covariance like FOIL]} \\
&= \mathsf{Var}(X) + 2\mathsf{Cov}(X, Y) + \mathsf{Var}(Y) && \text{[covariance with self, and symmetry]}
\end{aligned}$$

□

Example(s)

Let $X$ and $Y$ be two independent $\mathcal{N}(0, 1)$ random variables and:

$$Z = 1 + X + XY^2$$

$$W = 1 + X$$

Find $\text{Cov}(Z, W)$.

*Solution* First note that $\mathbb{E}\left[X^2\right] = \mathsf{Var}\left(X\right) + \mathbb{E}\left[X\right]^2 = 1 + 0^2 = 1$ (rearrange variance formula and solve for $\mathbb{E}\left[X^2\right]$). Similarly, $\mathbb{E}\left[Y^2\right] = 1$.

$$\begin{aligned}
\mathsf{Cov}\left(Z, W\right) &= \mathsf{Cov}\left(1 + X + XY^2, 1 + X\right) \\
&= \mathsf{Cov}\left(X + XY^2, X\right) && \text{[Property 4]} \\
&= \mathsf{Cov}\left(X, X\right) + \mathsf{Cov}\left(XY^2, X\right) && \text{[Property 7]} \\
&= \mathsf{Var}\left(X\right) + \mathbb{E}\left[X^2Y^2\right] - \mathbb{E}\left[XY^2\right]\mathbb{E}\left[X\right] && \text{[Property 2 and def of covariance]} \\
&= 1 + \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right] - \mathbb{E}\left[X\right]^2\mathbb{E}\left[Y^2\right] && \text{[Because } X \text{ and } Y \text{ are independent]} \\
&= 1 + 1 - 0 = 2
\end{aligned}$$

□

## 5.4.2 (Pearson) Correlation

Covariance has a "problem" in measuring linear relationships, in that $\mathsf{Cov}\left(X, Y\right)$ will be positive when there is a positive linear relationship and negative when there is a negative linear relationship, but $\mathsf{Cov}\left(2X, Y\right) = 2\mathsf{Cov}\left(X, Y\right)$. Scaling one of the random variables should not affect the *strength* of their relationship, which it seems to do. It would be great if we defined some metric that was normalized (had a maximum and minimum), and was invariant to scale. This metric will be called correlation!

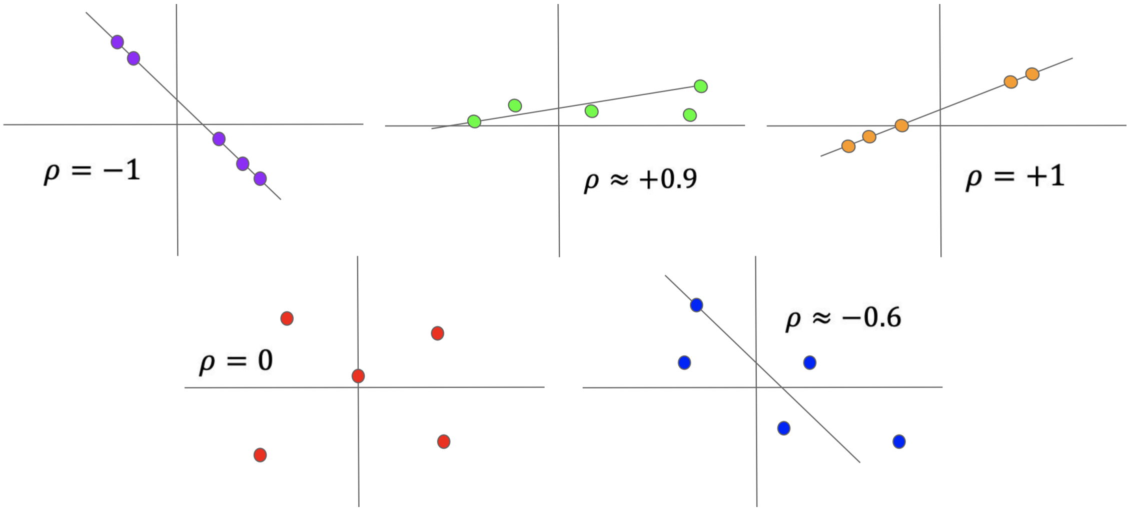
**Definition 5.4.2: (Pearson) Correlation**

Let $X, Y$ be random variables. The (Pearson) correlation of $X$ and $Y$ is:

$$\rho(X, Y) = \frac{\mathsf{Cov}\left(X, Y\right)}{\sqrt{\mathsf{Var}\left(X\right)}\sqrt{\mathsf{Var}\left(Y\right)}}$$

We can prove by the Cauchy-Schwarz inequality (from linear algebra), $-1 \leq \rho(X, Y) \leq 1$. That is, correlation is just a normalized version of covariance. Most notably, $\rho(X, Y) = \pm 1$ if and only if $Y = aX + b$ for some constants $a, b \in \mathbb{R}$, and then the sign of $\rho$ is the same as that of $a$.
In linear regression ("line-fitting") from high school science class, you may have calculated some $R^2$, $0 \leq R^2 \leq 1$, and this is actually $\rho^2$, and measure how well a linear relationship exists between $X$ and $Y$. $R^2$ is the percentage of variance in $Y$ which can be explained by $X$.

Let's take a look at some example graphs which shows a sample of data and their (Pearson) correlations, to get some intuition.



The 1st (purple) plot has a perfect negative linear relationship and so the correlation is $-1$.
The 2nd (green) plot has an positive relationship, but it is not perfect, so the correlation is around $+0.9$.
The 3rd (orange) plot is a perfectly linear positive relationship, so the correlation is $+1$.
The 4th (red) plot appears to have data that is independent, so the correlation is 0.
The 5th (blue) plot has a negative trend that isn't strongly linear, so the correlation is around $-0.6$.

Example(s)

Suppose $X$ and $Y$ are random variables, where $Y = -5X + 2$. Show that, since there is a perfect negative linear relationship, $\rho(X, Y) = -1$.

*Solution* To find the correlation, we need the covariance and the two individual variances. Let's write them in terms of $\mathsf{Var}(X)$.

$$\mathsf{Var}(Y) = \mathsf{Var}(-5X + 2) = (-5)^2\mathsf{Var}(X) = 25\mathsf{Var}(X)$$

By properties of covariance (shifting by 2 doesn't matter),

$$\mathsf{Cov}(X, Y) = \mathsf{Cov}(X, -5X + 2) = -5\mathsf{Cov}(X, X) = -5\mathsf{Var}(X)$$

Finally,

$$\rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}} = \frac{-5\mathsf{Var}(X)}{\sqrt{\mathsf{Var}(X)}\sqrt{25\mathsf{Var}(X)}} = \frac{-5\mathsf{Var}(X)}{5\mathsf{Var}(X)} = -1$$

Note that the $-5$ and 2 did not matter at all (except that $-5$ was negative and made the correlation negative)!
□

### 5.4.3 Variance of Sums of Random Variables

Perhaps the most useful application of covariance is in finding the variance of a sum of *dependent* random variables. We'll extend the case of $\mathsf{Var}(X + Y)$ to more than two random variables.

**Theorem 5.4.1: Variance of Sums of RVs**

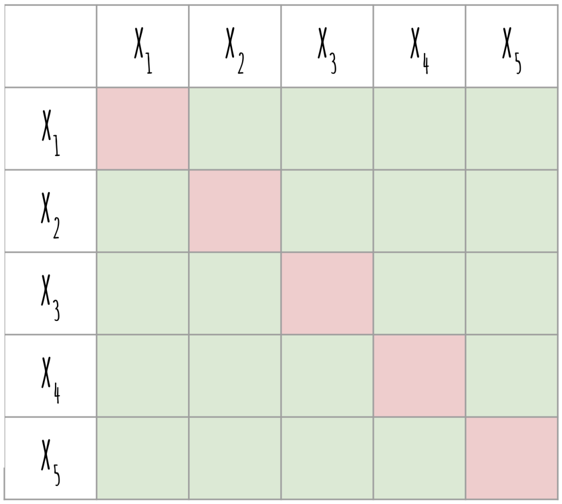
If $X_1, X_2, \dots, X_n$ are random variables, then

$$\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}\left(X_i\right) + 2\sum_{i<j} \mathsf{Cov}\left(X_i, X_j\right)$$

*Proof of Variance of Sums of RVs.* We'll first do something unintutive - making our expression more complicated. The variance of the sum $X_1 + X_2 + \cdots + X_n$ is the covariance with itself! We'll use $i$ to index one of the sums $\sum_{i=1}^{n} X_i$ and $j$ for the other $\sum_{j=1}^{n} X_i$. Keep in mind these both represent the same quantity; you'll see why we used different dummy variables soon!

$$\begin{aligned}
\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) && \text{[covariance with self = variance]} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{Cov}\left(X_i, X_j\right) && \text{[by FOIL]} \\
&= \sum_{i=1}^{n} \mathsf{Var}\left(X_i\right) + 2\sum_{i<j} \mathsf{Cov}\left(X_i, X_j\right) && \text{[by symmetry (see image below)]}
\end{aligned}$$

The final step comes from the definition of covariance of a variable with itself and the symmetry of the covariance. It is illustrated below where the red diagonal is the covariance of a variable with itself (which is its variance), and the green off-diagonal are the symmetric pairs of covariance. We used the fact that $\mathsf{Cov}\left(X_i, X_j\right) = \mathsf{Cov}\left(X_j, X_i\right)$ to require us to only sum the lower triangle (where $i < j$), and multiply by 2 to account for the upper triangle.



|  | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 |
| $X_2$ | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 |
| $X_3$ | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 |
| $X_4$ | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 |
| $X_5$ | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 |

It is important to remember than if all the RVs were independent, all the $\mathsf{Cov}\left(X_i, X_j\right)$ terms (for $i \neq j$) would be zero, and so we would just be left with the sum of the variances as we showed earlier! □

Example(s)

Recall in the hat check problem in 3.3, we had $n$ people who go to a party and leave their hats with a hat check person. At the end of the party, the hats are returned randomly though.

We let $X$ be the number of people who get their original hat back. We solved for $\mathbb{E}[X]$ with indicator random variables $X_1, \ldots X_n$ for whether the $i$-th person got their hat back.

We showed that:

$$\begin{aligned}\mathbb{E}[X_i] &= \mathbb{P}(X_i = 1) \\ &= \mathbb{P}\left(i^{\text{th}} \text{ person get their hat back}\right) \\ &= \frac{1}{n}\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} \frac{1}{n} \\ &= n \cdot \frac{1}{n} \\ &= 1\end{aligned}$$

Above was all review: now compute $\mathsf{Var}(X)$.

*Solution* Recall that each $X_i \sim \text{Ber}\left(\frac{1}{n}\right)$ (1 with probability $\frac{1}{n}$, and 0 otherwise). (Remember these were NOT independent RVs, but we still could apply linearity of expectation.) In our previous proof, we showed that

$$\mathsf{Var}(X) = \mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}(X_i) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j)$$

Recall that $X_i, X_j$ are indicator random variables which are in $\{0,1\}$, so their product $X_iX_j \in \{0,1\}$ as well.

This allows us to calculate:

$$\begin{aligned}\mathbb{E}[X_iX_j] &= \mathbb{P}(X_iX_j = 1) && \text{[since indicator, is just probability of being 1]} \\ &= \mathbb{P}(X_i = 1, X_j = 1) && \text{[product is 1 if and only if both are 1]} \\ &= \mathbb{P}(X_i = 1)\,\mathbb{P}(X_j = 1 \mid X_i = 1) && \text{[chain rule]} \\ &= \frac{1}{n}\left(\frac{1}{n-1}\right)\end{aligned}$$

This is because we need both person $i$ and person $j$ to get their hat back: person $i$ gets theirs back with probability $\frac{1}{n}$, and *given* this is true, person $j$ gets theirs back with probability $\frac{1}{n-1}$

So, by definition of covariance (recall each $\mathbb{E}[X_i] = \frac{1}{n}$):

$$\begin{aligned}
\mathsf{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\
&= \frac{1}{n}\left(\frac{1}{n-1}\right) - \frac{1}{n}\cdot\frac{1}{n} && \text{[plug in]} \\
&= \frac{n}{n^2(n-1)} - \frac{n-1}{n^2(n-1)} && \text{[algebra]} \\
&= \frac{1}{n^2(n-1)} && \text{[algebra]}
\end{aligned}$$

Further, since $X_i$ is a Bernoulli (indicator) random variable:

$$\mathsf{Var}(X_i) = p(1-p) = \left(\frac{1}{n}\right)\left(1 - \frac{1}{n}\right)$$

Finally, we have

$$\begin{aligned}
\mathsf{Var}(X) &= \sum_{i=1}^{n} \mathsf{Var}(X_i) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j) && \text{[formula for variance of sum]} \\
&= \sum_{i=1}^{n} \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2\sum_{i<j} \frac{1}{n^2(n-1)} && \text{[plug in]} \\
&= n\left(\frac{1}{n}\right)\left(1 - \frac{1}{n}\right) + 2\binom{n}{2}\left(\frac{1}{n^2(n-1)}\right) && \text{[there are } \tbinom{n}{2} \text{ pairs with } i<j\text{]} \\
&= \left(1 - \frac{1}{n}\right) + 2\frac{n(n-1)}{2}\left(\frac{1}{n^2(n-1)}\right) \\
&= \left(1 - \frac{1}{n}\right) + \frac{1}{n} \\
&= 1
\end{aligned}$$

How many pairs are their with $i < j$? This is just $\binom{n}{2} = \frac{n(n-1)}{2}$ since we just choose two different elements. Another way to see this is that there was an $n \times n$ square, and we removed the diagonal of $n$ elements, so we are left with $n^2 - n = n(n-1)$. Divide by two to get just the lower half.
This is very surprising and interesting! When returning $n$ hats randomly and uniformly, the expected number of people who get their hat back is 1, and so is the variance! These don't even depend on $n$ at all! □ It takes practice to get used to these formula, so let's do one more problem.

**Example(s)**

Suppose we throw 12 balls independently and uniformly into 7 bins. What are the mean and variance of the number of empty bins after this process? (Hint: Indicators).

*Solution* Let $X$ be the total number of empty bins, and $X_1, \dots, X_7$ be the indicator of whether or not bin $i$ is empty so that $X = \sum_{i=1}^{7} X_i$. Then,

$$\mathbb{P}(X_i = 1) = \left(\frac{6}{7}\right)^{12}$$

since we need to avoid this bin (with probability 6/7) 12 times independently. That is,

$$X_i \sim \text{Ber}\left(p = \left(\frac{6}{7}\right)^{12}\right)$$

Hence, $\mathbb{E}\left[X_i\right] = p \approx 0.1573$ and $\mathsf{Var}\left(X_i\right) = p(1-p) \approx 0.1325$. These random variables are surely dependent, since knowing one bin is empty means the 12 balls had to go to the other 6 bins, making it less likely that another bin is empty.

However, dependence doesn't bother us for computing the expectation; by linearity of expectation, we get

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\sum_{i=1}^{7} X_i\right] = \sum_{i=1}^{7} \mathbb{E}\left[X_i\right] = \sum_{i=1}^{7} \left(\frac{6}{7}\right)^{12} = 7\left(\frac{6}{7}\right)^{12} \approx 1.1009$$

Now for the variance, we need to find $\mathsf{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]$ for $i \neq j$. Well, $X_i X_j \in \{0, 1\}$ since both $X_i, X_j \in \{0, 1\}$, so $X_i X_j$ is indicator/Bernoulli as well with

$$\mathbb{E}\left[X_i X_j\right] = \mathbb{P}\left(X_i X_j = 1\right) = \mathbb{P}\left(X_i = 1, X_j = 1\right) = \mathbb{P}\left(\text{both bin } i \text{ and } j \text{ are empty}\right) = \left(\frac{5}{7}\right)^{12}$$

since all the balls must go into the other 5 bins during each of the 12 independent throws. Finally,

$$\mathsf{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] = \left(\frac{5}{7}\right)^{12} - \left(\frac{6}{7}\right)^{12}\left(\frac{6}{7}\right)^{12} \approx -0.0071$$

Recall that $\mathsf{Var}\left(X_i\right) = p(1-p) \approx 0.1325$, and so putting this all together gives:

$$\begin{aligned}
\mathsf{Var}\left(X\right) &= \sum_{i=1}^{7} \mathsf{Var}\left(X_i\right) + 2\sum_{i<j} \mathsf{Cov}\left(X_i, X_j\right) && \text{[formula for variance of sum]} \\
&\approx \sum_{i=1}^{7} 0.1325 + 2\sum_{i<j}(-0.0071) && \text{[plug in approximate decimal values]} \\
&= 7 \cdot 0.1325 + 2\binom{7}{2}(-0.0071) \\
&\approx 0.62954
\end{aligned}$$

□

Recall the hypergeometric RV $X \sim \text{HypGeo}(N, K, n)$ which was the number of lollipops we get when we draw $n$ candies from a bag of $N$ total candies ($K \leq N$ are lollipops). We stated without proof that $\mathsf{Var}\left(X\right) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$. You have the tools now to prove this if you like using indicators and covariances, but we'll prove this later in 5.8 as well!