Covariance and Correlation

Miao Yu

Fei Tian College Middletown

Fall, 2024

Introduction to Covariance

Recap: Variance

- Variance measures the spread of a single random variable around its mean.

Introduction to Covariance

Recap: Variance

- Variance measures the spread of a single random variable around its mean.

What is Covariance?

- Covariance extends the concept of variance to two random variables.
- It measures the *linear relationship* between two variables, X and Y, by capturing how they vary together.
- A positive covariance means X and Y tend to increase or decrease together, while a negative covariance implies that one variable tends to increase as the other decreases.

Introduction to Covariance

Recap: Variance

- Variance measures the spread of a single random variable around its mean.

What is Covariance?

- Covariance extends the concept of variance to two random variables.
- It measures the *linear relationship* between two variables, X and Y, by capturing how they vary together.
- A positive covariance means X and Y tend to increase or decrease together, while a negative covariance implies that one variable tends to increase as the other decreases.

Why Learn Covariance?

- Covariance is crucial in understanding how variables interact in a dataset.
- It provides the foundation for more complex concepts like correlation and helps calculate the variance of sums of dependent variables.

Definition of Covariance

Definition: Covariance

Let X and Y be random variables. The covariance of X and Y is defined by:

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

Definition of Covariance

Definition: Covariance

Let X and Y be random variables. The covariance of X and Y is defined by:

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

Note: Covariance can be negative, unlike variance.

$E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

Derivation:

$$\begin{aligned} E[(X - E[X])(Y - E[Y])] &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

$E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

Derivation:

$$\begin{aligned} E[(X - E[X])(Y - E[Y])] &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Interpretation:

- This formula simplifies the covariance calculation by using $E[XY]$, the expectation of the product, minus $E[X]E[Y]$, the product of the expectations.
- If X and Y are independent, $E[XY] = E[X]E[Y]$, leading to $\text{Cov}(X, Y) = 0$.
- However, the reverse argument is not true. $\text{Cov}(X, Y) = 0$ does not necessarily indicate $X \perp Y$.

Intuition Behind Covariance

We start with the definition of covariance:

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

By LOTUS, this can be expressed as:

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

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$.

Intuition Behind Covariance

We start with the definition of covariance:

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

By LOTUS, this can be expressed as:

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

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Intuitively, we observe:

- $x > \mu_X, y > \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) > 0$
- $x < \mu_X, y < \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) > 0$
- $x < \mu_X, y > \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) < 0$
- $x > \mu_X, y < \mu_Y \Rightarrow (x - \mu_X)(y - \mu_Y) < 0$

Understanding the Sign of Covariance

- When X and Y are often both above or both below their means, covariance tends to be positive.
- When X is above its mean while Y is below (or vice versa), covariance tends to be negative.
- In summary, covariance is:
  - Positive if X and Y generally increase together.
  - Negative if an increase in X typically leads to a decrease in Y.

Graphic Illustration of Examples

Lab Notebook: Examples of Positive, Zero and Negative Covariance

Example: Covariance with Uniform(0,1) Variables

Setup: Let $X \sim \text{Unif}(0, 1)$ and $Y \sim \text{Unif}(0, 1)$.

We consider two cases:

Case 1: $Y = 1 - X$

Case 2: X **and** Y **are Independent**

Numerical simulations of both cases are given in the same notebook on the previous page.

Example: Negative Covariance with $Y = 1 - X$

Setup: Let $X \sim \text{Unif}(0, 1)$ and $Y \sim \text{Unif}(0, 1)$, $Y = 1 - X$.

Example: Negative Covariance with $Y = 1 - X$

Setup: Let $X \sim \text{Unif}(0,1)$ and $Y \sim \text{Unif}(0,1)$, $Y = 1 - X$.

Calculate $\mathbf{Cov}(X,Y)$:

$$E[X] = \frac{1}{2}, \quad E[Y] = E[1-X] = 1 - E[X] = \frac{1}{2}$$
$$E[XY] = E\left[X(1-X)\right] = E[X - X^2] = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$
$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12}$$

Example: Negative Covariance with $Y = 1 - X$

Setup: Let $X \sim \text{Unif}(0,1)$ and $Y \sim \text{Unif}(0,1)$, $Y = 1 - X$.

Calculate Cov(X, Y):

$$E[X] = \frac{1}{2}, \quad E[Y] = E[1 - X] = 1 - E[X] = \frac{1}{2}$$
$$E[XY] = E\left[X(1 - X)\right] = E[X - X^2] = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$
$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12}$$

Conclusion: $\text{Cov}(X, Y) = -\frac{1}{12}$, indicating a negative linear relationship.

Example: Zero Covariance with Independent X and Y

Case 2: X and Y are Independent

Example: Zero Covariance with Independent X and Y

Case 2: X and Y are Independent

Calculate Cov(X, Y):

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= \frac{1}{4} - \frac{1}{4} = 0\end{aligned}$$

Example: Zero Covariance with Independent X and Y

Case 2: X and Y are Independent

Calculate Cov(X, Y):

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= \frac{1}{4} - \frac{1}{4} = 0\end{aligned}$$

Conclusion: When X and Y are independent, their covariance is zero, as there is no linear relationship between them.

Properties of Covariance

Covariance satisfies the following properties:

- If $X \perp Y$, then $\text{Cov}(X, Y) = 0$.
- $\text{Cov}(X, X) = \text{Var}(X)$.
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
- $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$.
- $\text{Cov}(aX + bY, Z) = a \cdot \text{Cov}(X, Z) + b \cdot \text{Cov}(Y, Z)$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

Proof of Property 1: Covariance of Independent RVs is Zero

If X and Y are independent, then:

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

Consequently:

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

Thus, independence implies zero covariance, though the reverse is not necessarily true.

Proof of Property 6: Variance of Sum of RVs

For any random variables X and Y:

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

Expanding this using the linearity of covariance:

$$= \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Y)$$

Simplifying, we get:

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

Example: Covariance Calculation

Problem Statement

Let X and Y be two independent $N(0,1)$ random variables with:

$$Z = 1 + X + XY^2, \quad W = 1 + X$$

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

Solution: Step 1 - Expectations and Expanding Covariance

First, since X and Y are $N(0, 1)$, we know:

$$E[X] = 0, \quad E[Y] = 0, \quad E[X^2] = E[Y^2] = \text{Var}(X) + E[X]^2 = 1$$

Next, we expand $\text{Cov}(Z, W)$ using properties of covariance:

$$\text{Cov}(Z, W) = \text{Cov}(1 + X + XY^2, 1 + X)$$

By the linearity of covariance:

$$= \text{Cov}(X + XY^2, X) = \text{Cov}(X, X) + \text{Cov}(XY^2, X)$$

Solution: Step 2 - Calculating Each Term

For $\text{Cov}(X, X)$:

$$\text{Cov}(X, X) = \text{Var}(X) = 1$$

For $\text{Cov}(XY^2, X)$:

$$\text{Cov}(XY^2, X) = E[X^2Y^2] - E[X]E[Y^2]$$

Since X and Y are independent:

$$E[X^2Y^2] = E[X^2]E[Y^2] = 1 \cdot 1 = 1$$

Thus:

$$\text{Cov}(XY^2, X) = 1 - 0 = 1$$

Substituting the values back, we find:

$$\text{Cov}(Z, W) = \text{Var}(X) + \text{Cov}(XY^2, X) = 1 + 1 = 2$$

Covariance and Its Limitations

Recap: Covariance

- Covariance measures the linear relationship between two random variables X and Y.
- Positive covariance indicates a positive linear relationship, and negative covariance indicates a negative linear relationship.

Covariance and Its Limitations

Recap: Covariance

- Covariance measures the linear relationship between two random variables X and Y.
- Positive covariance indicates a positive linear relationship, and negative covariance indicates a negative linear relationship.

Limitations of Covariance

- Covariance depends on the scale of X and Y. For example, $\text{Cov}(2X, Y) = 2\text{Cov}(X, Y)$.
- Scaling one variable affects the covariance, which does not reflect the inherent "strength" of the relationship.
- We need a normalized metric that is invariant to scale.

Definition of Pearson Correlation

Definition: (Pearson) Correlation

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

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Definition of Pearson Correlation

Definition: (Pearson) Correlation

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

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Properties of Correlation

- $-1 \leq \rho(X, Y) \leq 1$ (can be shown using the Cauchy-Schwarz inequality).
- $\rho(X, Y) = \pm 1$ if and only if $Y = aX + b$ for constants $a, b \in \mathbb{R}$.
- The sign of $\rho(X, Y)$ matches the sign of a when $Y = aX + b$.

Interpretation of Correlation

- Correlation is a normalized version of covariance that measures the strength and direction of the *linear* relationship between two variables.
- Unlike covariance, correlation is unaffected by the scale of the variables.
- In statistics, ρ^2 (often denoted as R^2) represents the proportion of variance in Y that can be explained by X, commonly used in linear regression.

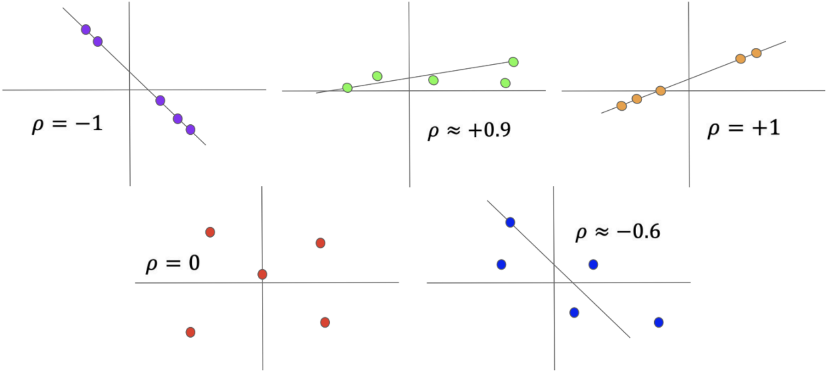
Interpretation of Correlation

- Correlation is a normalized version of covariance that measures the strength and direction of the *linear* relationship between two variables.
- Unlike covariance, correlation is unaffected by the scale of the variables.
- In statistics, ρ^2 (often denoted as R^2) represents the proportion of variance in Y that can be explained by X, commonly used in linear regression.

Summary

- $\rho = 1$: Perfect positive linear relationship.
- $\rho = -1$: Perfect negative linear relationship.
- $\rho = 0$: No linear relationship.

Graphic Examples of Correlation



Example Graphs and Intuition for Correlation

Consider the following example graphs with varying correlations:

- **Perfect Negative Correlation** ($\rho = -1$): A straight line with a negative slope.
- **Strong Positive Correlation** ($\rho \approx 0.9$): A positive trend but with some variability.
- **Perfect Positive Correlation** ($\rho = 1$): A straight line with a positive slope.
- **No Correlation** ($\rho = 0$): No apparent linear trend (variables are independent).
- **Weak Negative Correlation** ($\rho \approx -0.6$): A negative trend with more variability.

Example: Perfect Negative Correlation

Suppose $Y = -5X + 2$, where X and Y are random variables.

Step 1: Compute Variance of Y

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

Example: Perfect Negative Correlation

Suppose $Y = -5X + 2$, where X and Y are random variables.

Step 1: Compute Variance of Y

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

Step 2: Compute Covariance $\mathrm{Cov}(X, Y)$

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

Example: Computing Correlation $\rho(X, Y)$

Using the values from the previous slide:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Substitute in the values:

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

Example: Computing Correlation $\rho(X, Y)$

Using the values from the previous slide:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Substitute in the values:

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

Conclusion: Since $Y = -5X + 2$ represents a perfect negative linear relationship, the correlation $\rho(X, Y) = -1$, confirming the expected result.

Limitation of Pearson Correlation

Pearson Correlation Measures Only Linear Relationships

- The Pearson correlation coefficient $\rho(X, Y)$ measures the strength and direction of the *linear* relationship between two variables.
- If the relationship between X and Y is nonlinear, the Pearson correlation may be close to zero even if there is a strong association.

Limitation of Pearson Correlation

Pearson Correlation Measures Only Linear Relationships

- The Pearson correlation coefficient $\rho(X, Y)$ measures the strength and direction of the *linear* relationship between two variables.
- If the relationship between X and Y is nonlinear, the Pearson correlation may be close to zero even if there is a strong association.

Example of Nonlinear Relationship

- Consider $Y = X^2$ where $X \sim \text{Unif}(-1, 1)$.
- There is a clear relationship between X and Y, but it is nonlinear (a parabola).
- The Pearson correlation $\rho(X, Y)$ will be zero (a homework problem).

Limitation of Pearson Correlation

Pearson Correlation Measures Only Linear Relationships

- The Pearson correlation coefficient $\rho(X, Y)$ measures the strength and direction of the *linear* relationship between two variables.
- If the relationship between X and Y is nonlinear, the Pearson correlation may be close to zero even if there is a strong association.

Example of Nonlinear Relationship

- Consider $Y = X^2$ where $X \sim \text{Unif}(-1, 1)$.
- There is a clear relationship between X and Y, but it is nonlinear (a parabola).
- The Pearson correlation $\rho(X, Y)$ will be zero (a homework problem).

Conclusion: Pearson correlation is not suitable for detecting nonlinear relationships. Alternative measures like Spearman's rank correlation or mutual information can be used to capture nonlinear associations.

Variance of Sums of Random Variables

Motivation: Finding the variance of a sum of dependent random variables is an important application of covariance.

Variance of Sums of Random Variables

Motivation: Finding the variance of a sum of dependent random variables is an important application of covariance.

Objective: Extend the calculation of $\text{Var}(X+Y)$ to more than two random variables.

- For independent random variables $X_1, X_2, \ldots, X_n$, the variance of the sum simplifies to the sum of the variances.
- However, when variables are not independent, we must also account for their covariances.

Theorem: Variance of Sums of Random Variables

Theorem

Let $X_1, X_2, \dots, X_n$ be any random variables (without assuming independence). Then:

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

Theorem: Variance of Sums of Random Variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be any random variables (without assuming independence). Then:

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

Explanation:

- The variance of the sum includes the variances of each individual variable.
- The sum of covariances $\mathrm{Cov}(X_i, X_j)$ (for $i < j$) accounts for dependencies between the variables.

Proof Outline

Step 1: Express the Variance as a Covariance

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

Proof Outline

Step 1: Express the Variance as a Covariance

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

Step 2: Expand Using FOIL

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

Proof Outline

Step 1: Express the Variance as a Covariance

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

Step 2: Expand Using FOIL

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

Step 3: Separate the Diagonal and Off-Diagonal Terms

- The diagonal terms $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$.
- The off-diagonal terms $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$ contribute twice due to symmetry.

Result of the Proof

Conclusion:

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

Result of the Proof

Conclusion:

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

Interpretation:

- If all X_i are independent, $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, reducing the expression to $\sum_{i=1}^{n} \mathrm{Var}(X_i)$.
- If there are dependencies, the covariance terms adjust the total variance to reflect the interaction between variables.

Example: Variance of a Dependent Sum

Problem Statement

Problem: Recall the hat check problem with n people, where each person randomly receives a hat at the end.

Let X be the number of people who get their own hat back. Define indicator variables $X_1, X_2, \ldots, X_n$, where $X_i = 1$ if person i gets their hat back and 0 otherwise.

Goal: Compute $\mathrm{Var}(X)$.

Solution: Variance of the Sum in the Hat Check Problem

Step 1: Express Var(X) as a Sum of Variances and Covariances

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) = n\mathrm{Var}(X_1) + 2\binom{n}{2}\mathrm{Cov}(X_1, X_2)$$

since all $\mathrm{Var}(X_i)$ or $\mathrm{Cov}(X_i, X_j)$ are equal.

Solution: Variance of the Sum in the Hat Check Problem

Step 1: Express Var(X) as a Sum of Variances and Covariances

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) = n\mathrm{Var}(X_1) + 2\binom{n}{2}\mathrm{Cov}(X_1, X_2)$$

since all $\mathrm{Var}(X_i)$ or $\mathrm{Cov}(X_i, X_j)$ are equal.

Step 2: Calculate Var(X_i) for Each X_i

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

where $p = \frac{1}{n}$ is the probability that any given person receives their own hat.

Solution: Covariance in the Hat Check Problem

Step 3: Calculate Cov(X_i, X_j) **for** $i \neq j$

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$$

where:

$$E[X_i X_j] = \frac{1}{n} \cdot \frac{1}{n-1} = \frac{1}{n(n-1)}$$

and $E[X_i] = \frac{1}{n}$, so

$$\text{Cov}(X_i, X_j) = \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{1}{n^2(n-1)}$$

Final Calculation of $\mathrm{Var}(X)$

Step 4: Substitute the Variance and Covariance Terms

$$\mathrm{Var}(X) = n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2 \cdot \frac{n(n-1)}{2} \cdot \frac{1}{n^2(n-1)}$$

Simplifying, we get:

$$\mathrm{Var}(X) = 1 - \frac{1}{n} + \frac{1}{n} = 1$$

Final Calculation of Var(X)

Step 4: Substitute the Variance and Covariance Terms

$$\mathrm{Var}(X) = n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2 \cdot \frac{n(n-1)}{2} \cdot \frac{1}{n^2(n-1)}$$

Simplifying, we get:

$$\mathrm{Var}(X) = 1 - \frac{1}{n} + \frac{1}{n} = 1$$

Conclusion: In the hat check problem, the variance $\mathrm{Var}(X) = 1$, meaning that the variability of the number of people who receive their own hat remains constant, regardless of n.