Variance and covariance

Last modified: 2026-11-20 00:10:36 (PST)

Deviation, error, and noise

Definition 0.1 (Deviation)**.** A **deviation** is the difference between a value and a reference value. For any quantity z and reference value r, the deviation of z from r is:

$$z - r$$

Remark

Remark. In probability and statistics, "deviation" usually means deviation from a population mean.
See: Wikipedia: Deviation (statistics)

Example 0.1 (Deviation from a reference weight)**.** A newborn weighing 3,200 g deviates from a reference weight of 3,500 g by $3200 - 3500 = -300$ g.

Definition 0.2 (Deviation from a population or subpopulation mean (error, noise term))**.** The **deviation from the mean** of a random variable Y is its deviation from its expectation, also called its **error** or **noise term**:

$$\mathbb{d}(Y) \stackrel{\text{def}}{=} Y - \mathbb{E}[Y]$$

For a realized observation y:

$$\mathbb{d}(y) \stackrel{\text{def}}{=} y - \mathbb{E}[Y]$$

Remark

Remark. Other sources often call this quantity an **error** or **noise term** [1]. In regression settings, the reference mean is often conditional on covariates: $\mathbb{d}(y_i) \overset{\text{def}}{=} y_i - \mathbb{E}[Y_i \mid X_i = x_i]$.
These notes prefer "deviation" for the quantity just defined. The Morrison Lab's regression notes use "residual" (defined in their Linear regression chapter) for deviations from fitted values, write $\mathbb{d}(\cdot)$ for these model/data deviations, and reserve $\varepsilon(\cdot)$ for estimator-to-estimand deviations (see Estimation).
See:

- Wikipedia: Errors and residuals
- Wikipedia: Deviation (statistics)
- Wikipedia: Linear regression — Notation and terminology

Example 0.2 (Deviation of a die roll from its mean)**.** A fair die roll Y has $\mathbb{E}[Y] = 3.5$ (computed on the expectation page), so a roll of $y = 5$ has deviation

$$\begin{aligned}\mathbb{d}(5) &= 5 - 3.5 \\ &= 1.5,\end{aligned}$$

and a roll of $y = 2$ has deviation

$$\begin{aligned}\mathbb{d}(2) &= 2 - 3.5 \\ &= -1.5.\end{aligned}$$

Variance and related characteristics

Definition 0.1 (Variance)**.** The **variance** of a random variable X with $\mathbb{E}[X^2] < \infty$ is the expectation of the squared deviation from the mean [2]; that is:

$$\operatorname{Var}(X) \overset{\text{def}}{=} \mathbb{E}[[\mathbb{d}(X)]^2]$$

Theorem 0.1 (Variance as expected squared deviation from the mean)**.**

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

Proof

Proof. Substituting the definition of $\mathbb{d}(X)$ from Definition 0.2 into Definition 0.1:

$$\mathrm{Var}(X) \overset{\text{def}}{=} \mathbb{E}\big[[\mathbb{d}(X)]^2\big] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

□

Remark

Remark. The expanded form is the more commonly seen definition of the variance, and the canonical starting point in most treatments; it's given here as a theorem rather than the primary definition to keep $\mathbb{d}(X)$ as the definition's single building block, matching the notation of Definition 0.2.

Theorem 0.2 (Simplified expression for variance)**.**

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

Proof

Proof. Let $\mu = \mathbb{E}[X]$ be the population mean of X. It is a constant, so linearity of expectation applies with μ as a coefficient:

$$\begin{aligned}
\mathrm{Var}(X) &\overset{\text{def}}{=} \mathbb{E}\big[[\mathbb{d}(X)]^2\big] && \text{(definition of variance)} \\
&= \mathbb{E}\big[(X-\mu)^2\big] && \text{(definition of deviation from the population mean } \mu\text{)} \\
&= \mathbb{E}\big[X^2 - 2X\mu + (\mu)^2\big] && \text{(expand binomial square)} \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + (\mu)^2 && \text{(linearity of expectation; } \mu \text{ is a constant)} \\
&= \mathbb{E}[X^2] - 2(\mu)^2 + (\mu)^2 && (\mathbb{E}[X] = \mu) \\
&= \mathbb{E}[X^2] - (\mu)^2 && \text{(combine like terms)} \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && (\mu = \mathbb{E}[X])
\end{aligned}$$

□

Example 0.1 (Variance of a Bernoulli random variable)**.** Let $X \sim \mathrm{Ber}(\pi)$. Then $\mathbb{E}[X] = \pi$ and $\mathbb{E}[X^2] = \pi$ (both computed on the expectation page), so:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && \text{(simplified expression for variance)} \\
&= \pi - \pi^2 && \text{(substitute)} \\
&= \pi(1-\pi) && \text{(factor)}
\end{aligned}$$

For a fair coin ($\pi = 1/2$), $\mathrm{Var}(X) = 1/4$. Figure 1 graphs $\mathrm{Var}(X)$ against π.

Definition 0.2 (Conditional variance)**.** The **conditional variance** of Y given $X = x$ is the variance of Y under its conditional distribution given $X = x$:

$$\mathrm{Var}(Y \mid X = x) \stackrel{\text{def}}{=} \mathbb{E}[(Y - \mathbb{E}[Y \mid X = x])^2 \mid X = x]$$

Evaluating this function of x at X gives the random variable $\mathrm{Var}(Y \mid X)$, just as the conditional expectation function $\mathbb{E}[Y \mid X]$ comes from $\mathbb{E}[Y \mid X = x]$.

Example 0.2 (Conditional variance of a binary outcome)**.** Let (X, Y) have the joint PMF $\mathbb{P}(X = 0, Y = 0) = 0.2$, $\mathbb{P}(X = 0, Y = 1) = 0.3$, $\mathbb{P}(X = 1, Y = 0) = 0.1$, $\mathbb{P}(X = 1, Y = 1) = 0.4$ (the table in the expectation page's exercise). Given $X = 0$, Y is binary with

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = 0) &= 0.3/0.5 \\
&= 0.6,
\end{aligned}$$

so $\mathbb{E}[Y \mid X = 0] = 0.6$, and by Definition 0.2:

$$\begin{aligned}
\mathrm{Var}(Y \mid X = 0) &= (0 - 0.6)^2 \cdot 0.4 + (1 - 0.6)^2 \cdot 0.6 && \text{(definition of conditional variance)} \\
&= 0.144 + 0.096 && \text{(multiply)} \\
&= 0.24 && \text{(add)}
\end{aligned}$$

The same steps with

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = 1) &= 0.4/0.5 \\
&= 0.8
\end{aligned}$$

give

$$\begin{aligned}
\mathrm{Var}(Y \mid X = 1) &= (0.2)^2 \cdot 0.8 + (0.8)^2 \cdot 0.2 \\
&= 0.16.
\end{aligned}$$

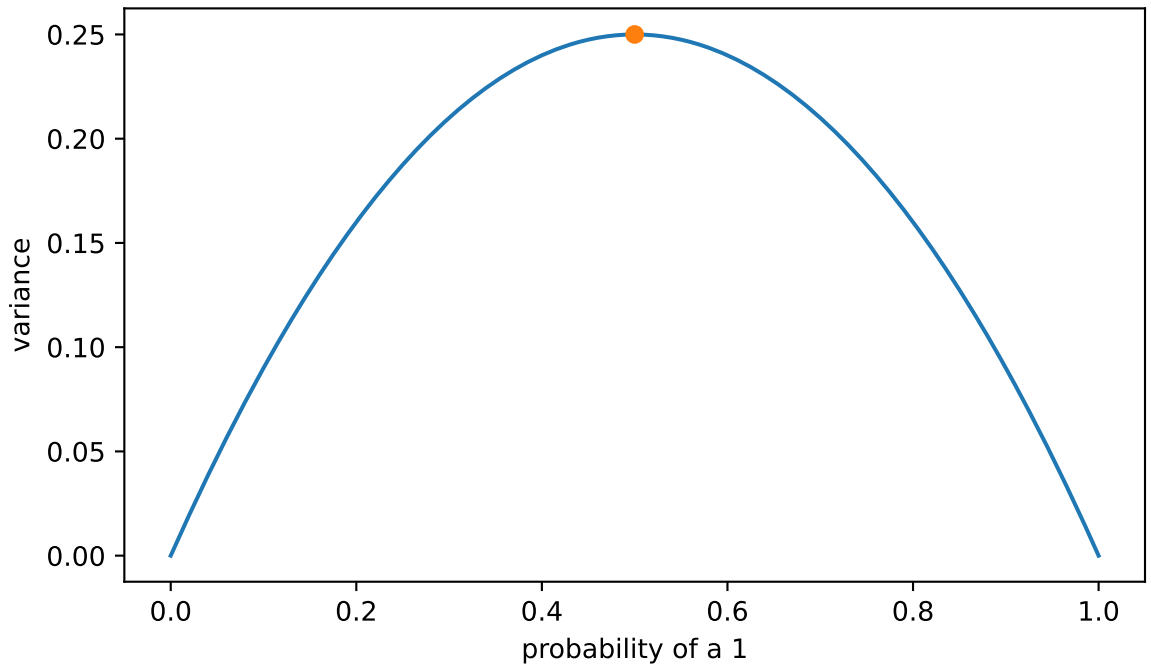
Definition 0.3 (Homoskedasticity)**.** A random variable Y is **homoskedastic** with respect to a random variable (or vector) X if the conditional variance of Y given $X = x$ is

Python

```python
import numpy as np
import matplotlib.pyplot as plt

prob = np.linspace(0, 1, 201)
variance = prob * (1 - prob)

fig, ax = plt.subplots(figsize=(6, 3.5), layout="constrained")
ax.plot(prob, variance, color="#1f77b4")
ax.plot(0.5, 0.5 * (1 - 0.5), "o", color="#ff7f0e")
ax.set_xlabel("probability of a 1")
ax.set_ylabel("variance")
plt.show()
```

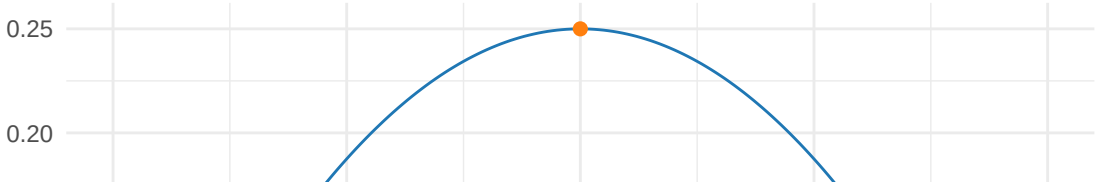


R

```r
library(ggplot2)

prob <- seq(0, 1, length.out = 201)
bernoulli <- data.frame(prob = prob, variance = prob * (1 - prob))
mid <- data.frame(prob = 0.5, variance = 0.5 * (1 - 0.5))

ggplot(bernoulli, aes(x = prob, y = variance)) +
  geom_line(colour = "#1f77b4") +
  geom_point(data = mid, colour = "#ff7f0e", size = 2) +
  labs(x = "probability of a 1", y = "variance") +
  theme_minimal()
```

the same for every x:

$$\mathrm{Var}(Y \mid X = x) = \sigma^2 \quad \text{for all } x$$

for some constant σ^2.

Example 0.3 (A homoskedastic total)**.** Flip two fair coins, let X_1 indicate heads on the first, and let Y be the total number of heads. Given $X_1 = 0$, Y is 0 or 1 with probability $1/2$ each; given $X_1 = 1$, Y is 1 or 2 with probability $1/2$ each. Either way, Y is $1/2$ away from its conditional mean with probability 1, so

$$\begin{aligned}\mathrm{Var}(Y \mid X_1 = 0) &= \mathrm{Var}(Y \mid X_1 = 1) \\ &= 1/4,\end{aligned}$$

and Y is homoskedastic with respect to X_1, with $\sigma^2 = 1/4$.

Definition 0.4 (Heteroskedasticity)**.** A random variable Y is **heteroskedastic** with respect to X if it is not homoskedastic with respect to X: that is, if $\mathrm{Var}(Y \mid X = x)$ differs between some values of x.

Example 0.4 (A heteroskedastic binary outcome)**.** In Example 0.2,

$$\begin{aligned}\mathrm{Var}(Y \mid X = 0) &= 0.24 \\ &\neq 0.16 \\ &= \mathrm{Var}(Y \mid X = 1),\end{aligned}$$

so Y is heteroskedastic with respect to X. More generally, a binary outcome with $\mathbb{P}(Y = 1 \mid X = x) = \pi(x)$ has conditional variance $\pi(x)(1 - \pi(x))$, so it is homoskedastic only if $\pi(x)$ takes at most two values, π and $1 - \pi$, for some π.

Theorem 0.3 (Law of total variance (conditional variance formula, Eve's law, variance decomposition formula))**.** *For random variables X and Y with $\mathbb{E}[Y^2] < \infty$:*

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])$$

Proof

Proof. Write $\mu \overset{\text{def}}{=} \mathbb{E}[Y]$ and $m(X) \overset{\text{def}}{=} \mathbb{E}[Y \mid X]$. Adding and subtracting $m(X)$ inside the deviation and expanding the square:

$$
\begin{aligned}
\operatorname{Var}(Y) &= \mathbb{E}\big[(Y-\mu)^2\big] && \text{(definition of variance)} \\
&= \mathbb{E}\big[([Y-m(X)]+[m(X)-\mu])^2\big] && \text{(add and subtract } m(X)\text{)} \\
&= \mathbb{E}\Big[(Y-m(X))^2\Big] + \mathbb{E}\Big[(m(X)-\mu)^2\Big] + 2\mathbb{E}[[Y-m(X)][m(X)-\mu]] && \text{(expand the square; linearity o}
\end{aligned}
$$

The three terms are handled separately, each using the law of iterated expectations, applied to a function of X and Y (conditional expectation of a function). The third term also uses two facts about conditional expectations: a function of X factors out of a conditional expectation given X, and conditional expectation is linear.
First term:

$$
\begin{aligned}
\mathbb{E}\Big[(Y-m(X))^2\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[(Y-m(X))^2 \mid X\Big]\Big] && \text{(law of iterated expectations)} \\
&= \mathbb{E}[\operatorname{Var}(Y \mid X)] && \text{(definition of conditional variance)}
\end{aligned}
$$

Second term: by the law of iterated expectations,

$$
\begin{aligned}
\mathbb{E}[m(X)] &= \mathbb{E}[\mathbb{E}[Y \mid X]] \\
&= \mu,
\end{aligned}
$$

so:

$$
\begin{aligned}
\mathbb{E}\Big[(m(X)-\mu)^2\Big] &= \mathbb{E}\Big[(m(X)-\mathbb{E}[m(X)])^2\Big] && (\mu = \mathbb{E}[m(X)]) \\
&= \operatorname{Var}(m(X)) && \text{(definition of variance)} \\
&= \operatorname{Var}(\mathbb{E}[Y \mid X]) && \text{(definition of } m(X)\text{)}
\end{aligned}
$$

Third term:

$$
\begin{aligned}
\mathbb{E}[[Y-m(X)][m(X)-\mu]] &= \mathbb{E}[\mathbb{E}[[Y-m(X)][m(X)-\mu] \mid X]] && \text{(law of iterated expectations)} \\
&= \mathbb{E}[[m(X)-\mu] \cdot \mathbb{E}[Y-m(X) \mid X]] && \text{(a function of } X \text{ factors out, with } g(X \\
&= \mathbb{E}[[m(X)-\mu] \cdot [\mathbb{E}[Y \mid X] - \mathbb{E}[m(X) \mid X]]] && \text{(linearity of conditional expectation)} \\
&= \mathbb{E}[[m(X)-\mu] \cdot [\mathbb{E}[Y \mid X] - m(X)]] && (\mathbb{E}[g(X) \mid X] = g(X), \text{ with } g = m) \\
&= \mathbb{E}[[m(X)-\mu] \cdot 0] && (\mathbb{E}[Y \mid X] = m(X)) \\
&= 0 && \text{(expectation of a constant)}
\end{aligned}
$$

Substituting the three terms:

$$
\begin{aligned}
\operatorname{Var}(Y) &= \mathbb{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\mathbb{E}[Y \mid X]) + 2 \cdot 0 && \text{(substitute the three terms)} \\
&= \mathbb{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\mathbb{E}[Y \mid X]) && \text{(simplify)}
\end{aligned}
$$

□

Example 0.5 (Decomposing the variance of a binary outcome)**.** Continuing Example 0.2,

$$\begin{aligned}\mathbb{P}(X=0) &= \mathbb{P}(X=1)\\ &= 0.5,\end{aligned}$$

$\mathbb{E}[Y \mid X=0] = 0.6$, and $\mathbb{E}[Y \mid X=1] = 0.8$, so $\mathbb{E}[\mathbb{E}[Y \mid X]] = 0.7$, and by Theorem 0.3:

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}(Y \mid X)] &= 0.5 \cdot 0.24 + 0.5 \cdot 0.16 && \text{(average the conditional variances)}\\
&= 0.20\\
\mathrm{Var}(\mathbb{E}[Y \mid X]) &= 0.5 \cdot (0.6-0.7)^2 + 0.5 \cdot (0.8-0.7)^2 && \text{(variance of the conditional means)}\\
&= 0.01\\
\mathrm{Var}(Y) &= 0.20 + 0.01 && \text{(law of total variance)}\\
&= 0.21
\end{aligned}$$

As a check, Y is Bernoulli with

$$\begin{aligned}\mathbb{P}(Y=1) &= 0.3 + 0.4\\ &= 0.7,\end{aligned}$$

so by Example 0.1,

$$\begin{aligned}\mathrm{Var}(Y) &= 0.7 \cdot 0.3\\ &= 0.21.\end{aligned}$$

Definition 0.5 (Precision)**.** The **precision** of a random variable X, often denoted $\tau(X)$, τ_X, or shorthanded as τ, is the inverse of that random variable's variance (for $\mathrm{Var}(X) > 0$); that is:

$$\tau(X) \stackrel{\text{def}}{=} (\mathrm{Var}(X))^{-1}$$

Definition 0.6 (Standard deviation)**.** The **standard deviation** of a random variable X is the square root of the variance of X:

$$\mathrm{SD}(X) \stackrel{\text{def}}{=} \sqrt{\mathrm{Var}(X)}$$

Remark

Remark. The standard deviation is on the same scale as X itself (unlike the variance, which is on the scale of X's square), which is why it is often the preferred measure of

spread when reporting results.

Example 0.6 (Precision and standard deviation of a fair coin flip)**.** In Example 0.1, a fair coin flip has $\mathrm{Var}(X) = 1/4$, so its precision is

$$\begin{aligned}\tau(X) &= 1/(1/4) \\ &= 4\end{aligned}$$

and its standard deviation is

$$\begin{aligned}\mathrm{SD}(X) &= \sqrt{1/4} \\ &= 1/2.\end{aligned}$$

Lemma 0.1 (The second moment is the variance plus the squared mean)**.** *For any random variable Z with $\mathbb{E}[Z^2] < \infty$:*

$$\mathbb{E}[Z^2] = \mathrm{Var}(Z) + (\mathbb{E}[Z])^2$$

Proof

Proof. Add $(\mathbb{E}[Z])^2$ to both sides of the simplified expression for variance, $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$. □

Example 0.7 (The second moment of a Bernoulli random variable)**.** Let Z be a Bernoulli random variable with success probability π, so $Z \sim \mathrm{Ber}(\pi)$. Then $\mathbb{E}[Z^2] = \pi$ directly, because $Z^2 = Z$ when Z is 0 or 1. The lemma gives the same value from the mean and the variance:

$$\begin{aligned}\mathrm{Var}(Z) + (\mathbb{E}[Z])^2 &= \pi(1-\pi) + (\pi)^2 && \text{(Bernoulli variance and mean)} \\ &= \pi && \text{(expand and combine like terms)}\end{aligned}$$

The best constant prediction

Exercise 0.1 (Predicting with one number)**.** A city must predict how many burglaries Y a randomly chosen neighborhood will have next year, using no input at all: it predicts the same number c for every neighborhood. Let μ be the mean $\mathbb{E}[Y]$ of Y. The expected squared error of c says how far that one number misses on average, so it is the quantity

the city wants to make small.
Express the expected squared error of the prediction c as a function of c.

Hint

Apply the squared error loss to the number c, and take its expectation.

Solution

Solution 0.1. The squared error of the prediction c is $(Y-c)^2$, so its expectation is $\mathbb{E}\left[(Y-c)^2\right]$.

Exercise 0.2 (Predicting with one number (part 2))**.** Again, Y is next year's burglary count in a randomly chosen neighborhood, c is a single number predicted for every neighborhood, and μ is the mean $\mathbb{E}[Y]$ of Y.
Split $\mathbb{E}\left[(Y-c)^2\right]$ into two parts: one that does not depend on c, and one that grows as c moves away from μ.

Hint

Inside the square, add and subtract μ. Then use the square of a sum, linearity of expectation, and the variance as an expected squared deviation.

Solution

Solution 0.2. Add and subtract μ inside the square:

$$\mathbb{E}\left[(Y-c)^2\right] = \mathbb{E}\left[((Y-\mu)+(\mu-c))^2\right].$$

Expand the square, using the square of a sum:

$$= \mathbb{E}\left[(Y-\mu)^2 + 2(Y-\mu)(\mu-c) + (\mu-c)^2\right].$$

Apply linearity of expectation. The numbers μ and c are constants, so $2(\mu-c)$ and $(\mu-c)^2$ come outside the expectation:

$$= \mathbb{E}\left[(Y-\mu)^2\right] + 2(\mu-c)\mathbb{E}[Y-\mu] + (\mu-c)^2.$$

The middle expectation is zero, because

$$\begin{aligned}\mathbb{E}[Y-\mu] &= \mathbb{E}[Y] - \mu \\ &= \mu - \mu \\ &= 0,\end{aligned}$$

so the squared-error risk reduces to:

$$= \mathbb{E}\left[(Y-\mu)^2\right] + (\mu - c)^2.$$

The first term is the variance of Y:

$$= \mathrm{Var}(Y) + (\mu - c)^2.$$

Theorem 0.1 (Expected squared error of a constant prediction)**.** *Let Y be a random variable with mean μ and finite variance. Then for every number c,*

$$\mathbb{E}\left[(Y-c)^2\right] = \mathrm{Var}(Y) + (\mu - c)^2. \tag{1}$$

Proof

Proof. Add and subtract μ inside the square, expand it, and take the expectation term by term:

$$\begin{aligned}
\mathbb{E}\left[(Y-c)^2\right] &= \mathbb{E}\left[((Y-\mu)+(\mu-c))^2\right] && \text{(add and subtract } \mu\text{)} \\
&= \mathbb{E}\left[(Y-\mu)^2 + 2(\mu-c)(Y-\mu) + (\mu-c)^2\right] && \text{(expand the square of a sum)} \\
&= \mathbb{E}\left[(Y-\mu)^2\right] + 2(\mu-c)\mathbb{E}[Y-\mu] + (\mu-c)^2 && \text{(linearity of expectation; } \mu \text{ and } c \text{ are constants)} \\
&= \mathrm{Var}(Y) + 2(\mu-c)\mathbb{E}[Y-\mu] + (\mu-c)^2 && \text{(variance as expected squared deviation)} \\
&= \mathrm{Var}(Y) + 2(\mu-c)\cdot 0 + (\mu-c)^2 && (\mathbb{E}[Y-\mu] = \mathbb{E}[Y]-\mu = 0) \\
&= \mathrm{Var}(Y) + (\mu-c)^2 && \text{(zero times a number)}
\end{aligned}$$

□

Exercise 0.3 (The best single number)**.** Use Theorem 0.1.
Which number c minimizes $\mathbb{E}\left[(Y-c)^2\right]$?

Hint

In Equation 1, ask which term depends on c, and when a square is smallest.

Solution

Solution 0.3. In Equation 1, the first term, $\mathrm{Var}(Y)$, does not depend on c. The second term, $(\mu-c)^2$, is a square, so it is never negative, and it is zero exactly when $c = \mu$. So

$$
\begin{aligned}
c &= \mu \\
&= \mathbb{E}[Y]
\end{aligned}
$$

minimizes the expected squared error, and no other number does. The set of minimizers has the single element μ.

Exercise 0.4 (The best single number (part 2))**.** Again, Y is next year's burglary count in a randomly chosen neighborhood, and c is a single number predicted for every neighborhood.
What is the smallest expected squared error that any constant prediction can achieve?

Hint

Use Equation 1 and your answer to Exercise 0.3.

Solution

Solution 0.4. At $c = \mu$, the second term is zero, so the smallest expected squared error is $\text{Var}(Y)$.
Under squared error, the best prediction made without any input is the mean, and what remains is the variance. A prediction that uses an input X can improve on $\text{Var}(Y)$ only through how the mean of Y changes with X. The conditional expectation function describes that change.

Corollary 0.1 (The mean is the best constant prediction under squared error)**.** *Under the assumptions of Theorem 0.1,*

$$
\arg\min_{c \in \mathbb{R}} \mathbb{E}\left[(Y-c)^2\right] = \{\mu\} \quad \textit{and} \quad \min_{c \in \mathbb{R}} \mathbb{E}\left[(Y-c)^2\right] = \text{Var}(Y),
$$

where arg min *is the set of minimizers.*

Proof

Proof. In Equation 1 the first term, $\text{Var}(Y)$, does not depend on c. The second term, $(\mu - c)^2$, is a square, so it is never negative, and it is zero exactly when $c = \mu$. So $c = \mu$ minimizes the expected squared error and no other number does, and at $c = \mu$ the expected squared error is $\text{Var}(Y)$. □

Example 0.1 (The best constant prediction of fuel economy)**.** Take the 392 cars in the Auto data described on the limit theorems page as the whole population, and predict

the fuel economy (`mpg`) of a car drawn at random with one number c. The code below computes the expected squared error of three constants, and sets each beside the right side of Equation 1.

```
auto <- readr::read_csv("data/auto.csv", show_col_types = FALSE)
mpg <- auto$mpg
mu <- mean(mpg)
squared_deviations <- (mpg - mu)^2
sigma2 <- mean(squared_deviations)

median_mpg <- median(mpg)
c_values <- c(mean = mu, median = median_mpg, twenty = 20)
constants <- names(c_values)

errors_by_car <- outer(mpg, c_values, "-")
squared_errors <- errors_by_car^2
errors <- colMeans(squared_errors)
shift_squared <- (mu - c_values)^2

tibble::tibble(
  constant = constants,
  c = c_values,
  expected_squared_error = errors,
  decomposition = sigma2 + shift_squared
)
#> # A tibble: 3 x 4
#>   constant     c expected_squared_error decomposition
#>   <chr>    <dbl>                 <dbl>         <dbl>
#> 1 mean      23.4                  60.8          60.8
#> 2 median    22.8                  61.2          61.2
#> 3 twenty    20                    72.6          72.6
```

Figure 2 plots the expected squared error of every constant from 10 to 40 against c, and marks the three constants in the table.

```r
c_grid <- seq(10, 40, length.out = 301)
grid_deviations <- outer(mpg, c_grid, "-")
grid_squared <- grid_deviations^2
grid_errors <- colMeans(grid_squared)
curve <- tibble::tibble(c = c_grid, expected_squared_error = grid_errors)
point_labels <- c("mean", "median", "c = 20")
label_hjust <- c(0.5, 0.5, 1)
label_vjust <- c(2.5, -2.5, -2)
marked <- tibble::tibble(
  point_label = point_labels,
  c = c_values,
  expected_squared_error = errors,
  label_hjust = label_hjust,
  label_vjust = label_vjust
)
curve_aes <- ggplot2::aes(x = c, y = expected_squared_error)
label_aes <- ggplot2::aes(
  label = point_label,
  hjust = label_hjust,
  vjust = label_vjust
)

curve |>
  ggplot2::ggplot() +
  curve_aes +
  ggplot2::geom_line() +
  ggplot2::geom_point(data = marked, size = 3) +
  ggplot2::geom_text(label_aes, data = marked) +
  ggplot2::labs(
    x = "constant prediction c (miles per gallon)",
    y = "expected squared error"
  ) +
  ggplot2::expand_limits(y = 0) +
  ggplot2::theme_minimal()
```

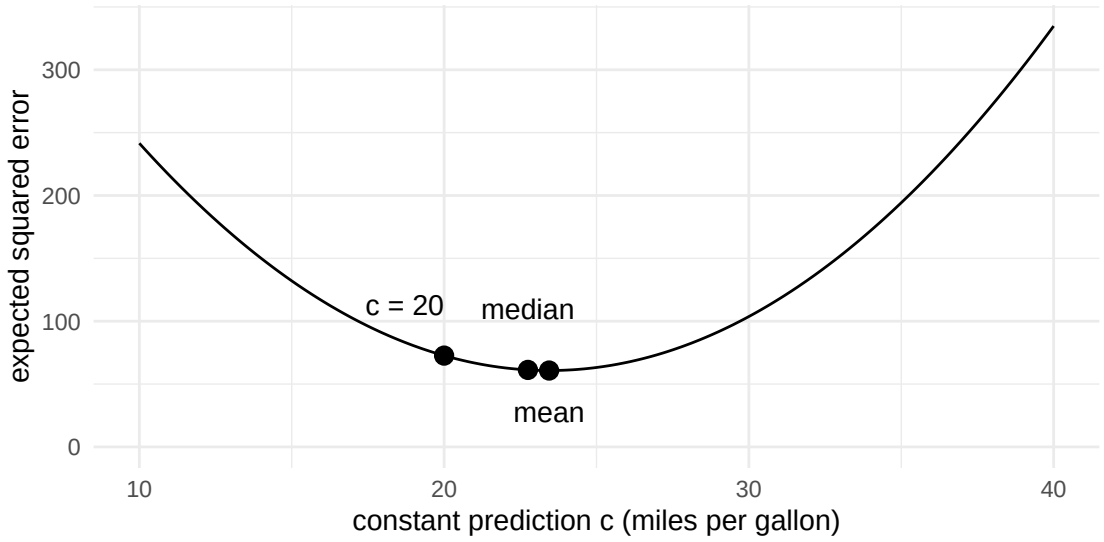


Figure 2: Expected squared error of a constant prediction c of fuel economy for the 392 cars in the Auto data, for c from 10 to 40 miles per gallon. The points mark the mean, the median and c = 20.

The two columns agree, as Theorem 0.1 says. The mean, 23.45, has the smallest expected squared error, 60.76, which is the variance of `mpg` over these cars (divisor the number of cars, not one fewer), as Corollary 0.1 says; the median and $c = 20$ are each worse by $(\mu - c)^2$.

Example 0.2 (Searching for the best constant)**.**
For the `mpg` values of the Auto data of Example 0.1, find the constant c that minimizes the average squared error by numerical search, without using the mean:

Python

```
import numpy as np
import pandas as pd
from scipy.optimize import minimize_scalar

mpg_search = pd.read_csv("data/auto.csv")["mpg"].to_numpy(dtype=float)


def average_squared_error(c):
    """The average squared error of the constant prediction c."""
    return np.mean((mpg_search - c) ** 2)


best_search = minimize_scalar(average_squared_error)
print(best_search.x, best_search.fun)
#> 23.445918367347897 60.76273844231571
print(mpg_search.mean(), mpg_search.var())
#> 23.445918367346938 60.76273844231571
```

R

```r
auto_search <- readr::read_csv("data/auto.csv", show_col_types = FALSE)
mpg_search <- auto_search$mpg
mpg_range <- range(mpg_search)
# The average squared error of the constant prediction c, minimized over c.
best_search <- optimize(
  \(c) {
    errors <- mpg_search - c
    mean(errors^2)
  },
  interval = mpg_range
)
mean_search <- mean(mpg_search)
deviations_search <- mpg_search - mean_search
variance_search <- mean(deviations_search^2)
c(
  minimizer = best_search$minimum,
  minimum = best_search$objective,
  mean = mean_search,
  variance = variance_search
)
#> minimizer   minimum      mean  variance
#>  23.44592  60.76274  23.44592  60.76274
```

The search ends at 23.446, the mean of `mpg`, and the smallest value it finds is 60.763, the variance of `mpg` over these cars. This matches Corollary 0.1: the best constant under squared error is the mean, and its expected squared error is the variance.

Independent noise plus a random variable

Mean-zero noise model

Measurement noise is almost always modeled as a random variable ε with **mean zero** and variance σ^2.

> **Remark (Mean zero is a parametrization choice)**
>
> *Remark* 0.1 (Mean zero is a parametrization choice). Mean zero is a parametrization choice, not a substantive assumption. In $Y = f(x) + \varepsilon$, moving a constant from ε into $f(x)$ changes the split but not the distribution of Y, so the data identify only the sum

$f(x) + \mathbb{E}[\varepsilon]$, the mean of Y at x. Requiring $\mathbb{E}[\varepsilon] = 0$ picks the one split in which $f(x)$ is that mean. This loses no generality when the model for f can absorb a constant, as any model with an intercept can. It restricts the model only when f cannot absorb one, as in a regression through the origin.

Mean zero gives one identity, used repeatedly across statistical modeling and machine learning.

Exercise 0.1 (The average squared error of pure noise)**.** Let ε be a random variable with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$.
Find $\mathbb{E}[\varepsilon^2]$ in terms of σ^2, and say which earlier result you used.

Solution

Solution 0.1. Theorem 0.2 says $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ for any X. Taking $X = \varepsilon$ and substituting $\mathbb{E}[\varepsilon] = 0$,

$$\begin{aligned} \mathbb{E}[\varepsilon^2] &= \mathrm{Var}(\varepsilon) + 0^2 \\ &= \sigma^2 \end{aligned} \tag{2}$$

so for mean-zero noise the average squared value *is* the variance.
The identity looks slight and does a lot of work. Equation 2 is the step that produces the residual variance term in regression and statistical modeling: it converts a statement about the *spread* of the noise, which is what a model of the world gives you, into a statement about *squared error*, which is what the fitting objective measures.

Theorem 0.1 (The second moment of mean-zero noise)**.** *Let ε be a random variable with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. Then*

$$\mathbb{E}[\varepsilon^2] = \sigma^2$$

Proof

Proof. This is the solution to Exercise 0.1, worked in Solution 0.1. It rearranges Theorem 0.2, using $\mathbb{E}[\varepsilon] = 0$. □

Theorem 0.2 (The second moment of independent noise plus a random variable)**.** *Let ε be a random variable with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. Let X be a random variable with $\mathbb{E}[X^2] < \infty$ that is independent of ε. Then*

$$\mathbb{E}\left[(\varepsilon + X)^2\right] = \sigma^2 + \operatorname{Var}(X) + \left(\mathbb{E}[X]\right)^2$$

Proof

Proof. Expanding the square and using linearity of expectation, first for the expectation of a sum and then for a constant factor,

$$\begin{aligned}
\mathbb{E}\left[(\varepsilon + X)^2\right] &= \mathbb{E}\left[\varepsilon^2 + 2\varepsilon X + X^2\right] && \text{(expand: } (a+b)^2 = a^2 + 2ab + b^2\text{)} \\
&= \mathbb{E}\left[\varepsilon^2\right] + \mathbb{E}[2\varepsilon X] + \mathbb{E}\left[X^2\right] && \text{(expectation of a sum)} \\
&= \mathbb{E}\left[\varepsilon^2\right] + 2\mathbb{E}[\varepsilon X] + \mathbb{E}\left[X^2\right] && \text{(constant factor 2)}
\end{aligned}$$

Each of the three terms is evaluated in turn.

- $\mathbb{E}\left[\varepsilon^2\right] = \sigma^2$, by Theorem 0.1.
- Because ε and X are independent, the expectation of the product is the product of the expectations:

$$\begin{aligned}
\mathbb{E}[\varepsilon X] &= \mathbb{E}[\varepsilon]\mathbb{E}[X] && \text{(independence)} \\
&= 0 \cdot \mathbb{E}[X] && (\mathbb{E}[\varepsilon] = 0) \\
&= 0 && \text{(zero times a number)}
\end{aligned}$$

- $\mathbb{E}\left[X^2\right] = \operatorname{Var}(X) + \left(\mathbb{E}[X]\right)^2$, by Theorem 0.2.

Substituting the three terms gives

$$\begin{aligned}
\mathbb{E}\left[(\varepsilon + X)^2\right] &= \sigma^2 + 2 \cdot 0 + \left(\operatorname{Var}(X) + \left(\mathbb{E}[X]\right)^2\right) && \text{(the three terms above)} \\
&= \sigma^2 + \operatorname{Var}(X) + \left(\mathbb{E}[X]\right)^2 && \text{(drop } 2 \cdot 0\text{)}
\end{aligned}$$

□

Remark (Use in statistics and machine learning)

Remark (Use in statistics and machine learning). Let ε be the noise in a new outcome Y, and let X be the difference between the true mean $\mu = \mathbb{E}[Y]$ and a prediction $\hat{y}$:

$$X \stackrel{\text{def}}{=} \mu - \hat{y}$$

The mean of X is minus the bias of the prediction:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[\mu - \hat{y}] && \text{(definition of } X\text{)} \\
&= \mu - \mathbb{E}[\hat{y}] && \text{(expectation of a difference; } \mu \text{ is a constant)} \\
&= -(\mathbb{E}[\hat{y}] - \mu) && \text{(factor out } -1\text{)} \\
&= -\text{Bias}(\hat{y}) && \text{(definition of bias)}
\end{aligned}$$

The variance of X is the variance of the prediction:

$$\begin{aligned}
\text{Var}(X) &= \text{Var}(\mu - \hat{y}) && \text{(definition of } X\text{)} \\
&= \text{Var}(\hat{y}) && \text{(a constant shifts a variable without changing its variance, and a sign flip does not c}
\end{aligned}$$

Substituting these two results into the theorem, with $(-\text{Bias}(\hat{y}))^2 = (\text{Bias}(\hat{y}))^2$, gives the squared prediction error of the new outcome as

$$\mathbb{E}\left[(\varepsilon + X)^2\right] = \sigma^2 + \text{Var}(\hat{y}) + (\text{Bias}(\hat{y}))^2,$$

where σ^2 is the residual variance. The statistics notes define the bias of a prediction and the training data it depends on, and state this result for predictions as the bias, variance and noise decomposition, with its exercise, worked example and definition of the residual variance.

Covariance

Definition 0.1 (Covariance)**.** The **covariance** of two random variables X and Y with $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$ [3] is:

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

Theorem 0.1 (Alternative formula for covariance)**.**

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

Proof

Proof. Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$ be the population means of X and Y. They are constants, so linearity of expectation applies with them as coefficients, analogous to the proof of the simplified expression for variance:

$$
\begin{aligned}
\operatorname{Cov}(X,Y) &\overset{\text{def}}{=} \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] && \text{(definition of covariance, with population means } \mu_X, \mu_Y\text{)} \\
&= \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] && \text{(expand the product)} \\
&= \mathbb{E}[XY] - \mu_Y\mathbb{E}[X] - \mu_X\mathbb{E}[Y] + \mu_X\mu_Y && \text{(linearity of expectation; } \mu_X, \mu_Y \text{ are constants)} \\
&= \mathbb{E}[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y && (\mathbb{E}[X] = \mu_X \text{ and } \mathbb{E}[Y] = \mu_Y) \\
&= \mathbb{E}[XY] - \mu_X\mu_Y && \text{(combine like terms)} \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && (\mu_X = \mathbb{E}[X] \text{ and } \mu_Y = \mathbb{E}[Y])
\end{aligned}
$$

□

Example 0.1 (Covariance of a binary exposure and outcome)**.** For the joint PMF in Example 0.2,

$$
\begin{aligned}
\mathbb{E}[XY] &= \mathbb{P}(X=1, Y=1) \\
&= 0.4,
\end{aligned}
$$

$\mathbb{E}[X] = 0.5$, and $\mathbb{E}[Y] = 0.7$, so:

$$
\begin{aligned}
\operatorname{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && \text{(alternative formula for covariance)} \\
&= 0.4 - 0.5 \cdot 0.7 && \text{(substitute)} \\
&= 0.05 && \text{(evaluate)}
\end{aligned}
$$

Theorem 0.2 (Independent random variables have zero covariance)**.** *If X and Y are independent, each discrete or continuous, with defined expectations, then:*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

If also $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$, then $\operatorname{Cov}(X,Y) = 0$.

Proof

Proof. Write f_X and f_Y for the PMFs or densities of X and Y, with reference measures μ_X and μ_Y as in the joint-distribution form of Fubini–Tonelli (counting measure for a discrete variable, Lebesgue measure for a continuous one). Because X and Y are independent, $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$ is their joint PMF, density, or density-mass function (the factorization in the notes to the definition of independence). First, with $h(x,y) = |x||y| \geq 0$ (condition (a)):

$$\begin{aligned}
\mathbb{E}[|XY|] &= \int_{\mathcal{R}(X)} \left(\int_{\mathcal{R}(Y)} |x||y|\, f_X(x)\, f_Y(y)\, d\mu_Y(y) \right) d\mu_X(x) && \text{(joint-distribution form of Fubini–Tonelli, condit} \\
&= \int_{\mathcal{R}(X)} |x|\, f_X(x) \left(\int_{\mathcal{R}(Y)} |y|\, f_Y(y)\, d\mu_Y(y) \right) d\mu_X(x) && (|x|\, f_X(x) \text{ does not depend on } y) \\
&= \int_{\mathcal{R}(X)} |x|\, f_X(x) \cdot \mathbb{E}[|Y|]\, d\mu_X(x) && \text{(LOTUS for } |Y|) \\
&= \mathbb{E}[|X|] \cdot \mathbb{E}[|Y|] && \text{(LOTUS for } |X|)
\end{aligned}$$

which is finite, because X and Y have defined expectations. So condition (b) holds for $h(x, y) = xy$, and the same steps without the absolute values give:

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{\mathcal{R}(X)} \left(\int_{\mathcal{R}(Y)} xy\, f_X(x)\, f_Y(y)\, d\mu_Y(y) \right) d\mu_X(x) && \text{(joint-distribution form of Fubini–Tonelli, condition} \\
&= \int_{\mathcal{R}(X)} x\, f_X(x) \left(\int_{\mathcal{R}(Y)} y\, f_Y(y)\, d\mu_Y(y) \right) d\mu_X(x) && (x\, f_X(x) \text{ does not depend on } y) \\
&= \int_{\mathcal{R}(X)} x\, f_X(x) \cdot \mathbb{E}[Y]\, d\mu_X(x) && \text{(definition of expectation)} \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y] && \text{(definition of expectation)}
\end{aligned}$$

Finally, by the alternative formula for covariance:

$$\begin{aligned}
\operatorname{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && \text{(alternative formula for covariance)} \\
&= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] && (\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]) \\
&= 0 && \text{(subtract)}
\end{aligned}$$

□

Example 0.2 (Two independent coin flips)**.** For the independent coin flips X_1 and X_2 of the independence page's example, $X_1 X_2 = 1$ only when both flips are heads, so:

$$\begin{aligned}
\mathbb{E}[X_1 X_2] &= \mathbb{P}(X_1 = 1, X_2 = 1) && (X_1 X_2 \text{ is the indicator of two heads}) \\
&= \tfrac{1}{4} && (\text{each outcome has probability } \tfrac{1}{4}) \\
&= \tfrac{1}{2} \cdot \tfrac{1}{2} && \text{(factor)} \\
&= \mathbb{E}[X_1]\mathbb{E}[X_2] && (\text{each flip is } \operatorname{Ber}(1/2))
\end{aligned}$$

so $\operatorname{Cov}(X_1, X_2) = 0$, as Theorem 0.2 requires.

Example 0.3 (Zero covariance without independence)**.** The converse of Theorem 0.2 is false. Let X take the values -1, 0, and 1 with probability $1/3$ each, and let $Y = X^2$. Then

$$\begin{aligned}\mathbb{E}[X] &= (-1+0+1)/3 \\ &= 0,\end{aligned}$$

and

$$\begin{aligned}XY &= X^3 \\ &= X,\end{aligned}$$

so:

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && \text{(alternative formula for covariance)} \\ &= \mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] && (XY = X^3 \text{ and } X^3 = X \text{ on } \{-1,0,1\}) \\ &= 0 - 0 \cdot \mathbb{E}[Y] && (\mathbb{E}[X] = 0) \\ &= 0 && \text{(multiply)}\end{aligned}$$

But X and Y are not independent: Y is a function of X, and

$$\begin{aligned}\mathbb{P}(X = 0, Y = 0) &= \mathbb{P}(X = 0) \\ &= \tfrac{1}{3},\end{aligned}$$

while

$$\begin{aligned}\mathbb{P}(X = 0)\,\mathbb{P}(Y = 0) &= \tfrac{1}{3} \cdot \tfrac{1}{3} \\ &= \tfrac{1}{9}.\end{aligned}$$

Definition 0.2 (Correlation)**.** The **correlation** of two random variables X and Y with finite, positive variances is their covariance divided by the product of their standard deviations:

$$\operatorname{Cor}(X,Y) \overset{\text{def}}{=} \frac{\operatorname{Cov}(X,Y)}{\operatorname{SD}(X)\operatorname{SD}(Y)}$$

i Remark

Remark. Dividing by the standard deviations removes the units of X and Y. This population correlation is a property of a joint distribution; the sample (Pearson) correlation coefficient computed from data estimates it.

Example 0.4 (Correlation of a binary exposure and outcome)**.** In Example 0.1, $\mathrm{Cov}(X, Y) = 0.05$, where $X \sim \mathrm{Ber}(0.5)$ has $\mathrm{Var}(X) = 0.25$ (Example 0.1) and Y has $\mathrm{Var}(Y) = 0.21$ (Example 0.5), so:

$$\begin{aligned}
\mathrm{Cor}(X, Y) &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} && \text{(definition of correlation)} \\
&= \frac{0.05}{\sqrt{0.25} \cdot \sqrt{0.21}} && \text{(substitute; } \mathrm{SD}(\cdot) = \sqrt{\mathrm{Var}(\cdot)}\text{)} \\
&\approx \frac{0.05}{0.5 \cdot 0.458} && \text{(evaluate the square roots)} \\
&\approx 0.218 && \text{(divide)}
\end{aligned}$$

Definition 0.3 (Uncorrelated random variables)**.** Random variables X and Y with $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$ are **uncorrelated** when their covariance is 0:

$$\mathrm{Cov}(X, Y) = 0$$

Remark

Remark. Theorem 0.2 says independent random variables are uncorrelated, and Example 0.3 shows the converse fails.

Example 0.5 (Correlated and uncorrelated pairs)**.** In Example 0.3, $\mathrm{Cov}(X, Y) = 0$, so X and $Y = X^2$ are uncorrelated. In Example 0.1, $\mathrm{Cov}(X, Y) = 0.05 \neq 0$, so the binary exposure and outcome are not uncorrelated.

Corollary 0.1 (Uncorrelated means zero correlation)**.** *If X and Y have finite, positive variances, then X and Y are uncorrelated if and only if* $\mathrm{Cor}(X, Y) = 0$.

Proof

Proof. By Definition 0.2, $\mathrm{Cor}(X, Y) = \mathrm{Cov}(X, Y)/(\mathrm{SD}(X)\,\mathrm{SD}(Y))$, and $\mathrm{SD}(X)\,\mathrm{SD}(Y) > 0$ because both variances are positive, so $\mathrm{Cor}(X, Y) = 0$ if and only if $\mathrm{Cov}(X, Y) = 0$, which is Definition 0.3. □

Definition 0.4 (Conditional covariance)**.** The **conditional covariance** of Y and Z given $X = x$ is their covariance under their conditional distribution given $X = x$:

$$\mathrm{Cov}(Y, Z \mid X = x) \overset{\text{def}}{=} \mathbb{E}[(Y - \mathbb{E}[Y \mid X = x])(Z - \mathbb{E}[Z \mid X = x]) \mid X = x]$$

Evaluating this function of x at X gives the random variable $\mathrm{Cov}(Y, Z \mid X)$.

Example 0.6 (Conditional covariance of a variable with itself)**.** Taking $Z = Y$ in Definition 0.4 gives the conditional variance: $\mathrm{Cov}(Y, Y \mid X = x) = \mathrm{Var}(Y \mid X = x)$. In Example 0.2, for example, $\mathrm{Cov}(Y, Y \mid X = 0) = 0.24$.

Theorem 0.3 (Law of total covariance (covariance decomposition formula, conditional covariance formula))**.** *For random variables X, Y, and Z with $\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[Z^2] < \infty$:*

$$\mathrm{Cov}(Y, Z) = \mathbb{E}[\mathrm{Cov}(Y, Z \mid X)] + \mathrm{Cov}(\mathbb{E}[Y \mid X], \mathbb{E}[Z \mid X])$$

Proof

Proof. The proof follows the proof of the law of total variance, which is the special case $Z = Y$. Write $m_Y(X) \overset{\text{def}}{=} \mathbb{E}[Y \mid X]$, $m_Z(X) \overset{\text{def}}{=} \mathbb{E}[Z \mid X]$, $\mu_Y \overset{\text{def}}{=} \mathbb{E}[Y]$, and $\mu_Z \overset{\text{def}}{=} \mathbb{E}[Z]$. Adding and subtracting $m_Y(X)$ and $m_Z(X)$ inside the deviations:

$$\begin{aligned}
\mathrm{Cov}(Y, Z) &= \mathbb{E}[(Y - \mu_Y)(Z - \mu_Z)] && \text{(definition of covariance)} \\
&= \mathbb{E}[([Y - m_Y(X)] + [m_Y(X) - \mu_Y])([Z - m_Z(X)] + [m_Z(X) - \mu_Z])] && \text{(add and subtract)} \\
&= \mathbb{E}[[Y - m_Y(X)][Z - m_Z(X)]] + \mathbb{E}[[Y - m_Y(X)][m_Z(X) - \mu_Z]] \\
&\quad + \mathbb{E}[[m_Y(X) - \mu_Y][Z - m_Z(X)]] + \mathbb{E}[[m_Y(X) - \mu_Y][m_Z(X) - \mu_Z]] && \text{(expand the product; linea}
\end{aligned}$$

The two middle terms are 0, by the same steps as the cross term in the proof of Theorem 0.3: condition on X by the law of iterated expectations, factor out the function of X (pull-out property), and use $\mathbb{E}[Y - m_Y(X) \mid X] = 0$ (or $\mathbb{E}[Z - m_Z(X) \mid X] = 0$), which follows from linearity of conditional expectation. For the first term:

$$\begin{aligned}
\mathbb{E}[[Y - m_Y(X)][Z - m_Z(X)]] &= \mathbb{E}[\mathbb{E}[[Y - m_Y(X)][Z - m_Z(X)] \mid X]] && \text{(law of iterated expectations)} \\
&= \mathbb{E}[\mathrm{Cov}(Y, Z \mid X)] && \text{(definition of conditional covariance)}
\end{aligned}$$

For the last term, the law of iterated expectations gives $\mathbb{E}[m_Y(X)] = \mu_Y$ and $\mathbb{E}[m_Z(X)] = \mu_Z$, so:

$$\begin{aligned}
\mathbb{E}[[m_Y(X) - \mu_Y][m_Z(X) - \mu_Z]] &= \mathbb{E}[[m_Y(X) - \mathbb{E}[m_Y(X)]][m_Z(X) - \mathbb{E}[m_Z(X)]]] && \text{(substitute the means)} \\
&= \mathrm{Cov}(m_Y(X), m_Z(X)) && \text{(definition of covariance)} \\
&= \mathrm{Cov}(\mathbb{E}[Y \mid X], \mathbb{E}[Z \mid X]) && \text{(definitions of } m_Y, m_Z\text{)}
\end{aligned}$$

Adding the four terms gives the result. □

Lemma 0.1 (The covariance of a variable with itself is its variance)**.** *For any random variable X with $\mathbb{E}[X^2] < \infty$:*

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

Proof

Proof. By the alternative formula for covariance:

$$\begin{aligned}
\text{Cov}(X, X) &= \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] && \text{(alternative formula for covariance, with } Y = X) \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && (XX = X^2) \\
&= \text{Var}(X) && \text{(simplified expression for variance)}
\end{aligned}$$

□

Definition 0.5 (Variance/covariance of a $p \times 1$ random vector)**.** For a $p \times 1$ dimensional random vector $\underset{\sim}{X}$,

$$\begin{aligned}
\text{Var}(\underset{\sim}{X}) &\stackrel{\text{def}}{=} \text{Cov}(\underset{\sim}{X}) \\
&\stackrel{\text{def}}{=} \mathbb{E}\left[(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})^\top\right]
\end{aligned}$$

Theorem 0.4 (Elements of the variance-covariance matrix are pairwise covariances)**.** *For a $p \times 1$ random vector $\underset{\sim}{X} = (X_1, \ldots, X_p)^\top$, the (i, j)-th element of $\text{Var}(\underset{\sim}{X})$ is $\text{Cov}(X_i, X_j)$:*

$$\text{Var}(\underset{\sim}{X}) = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}(X_p) \end{pmatrix}$$

Proof

Proof. Let $\mu_i = \mathbb{E}[X_i]$ for $i = 1, \ldots, p$, so $\mathbb{E}\underset{\sim}{X} = (\mu_1, \ldots, \mu_p)^\top$. By Definition 0.5:

$$
\begin{aligned}
\operatorname{Var}(\underset{\sim}{X}) &= \mathbb{E}\left[(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})^\top\right] \\
&= \mathbb{E}\left[\begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_p \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & \cdots & X_p - \mu_p \end{pmatrix}\right] \\
&= \mathbb{E}\left[\begin{pmatrix} (X_1 - \mu_1)(X_1 - \mu_1) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & \cdots & (X_p - \mu_p)(X_p - \mu_p) \end{pmatrix}\right] \\
&= \begin{pmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & \mathbb{E}\left[(X_1 - \mu_1)(X_p - \mu_p)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[(X_p - \mu_p)(X_1 - \mu_1)\right] & \cdots & \mathbb{E}\left[(X_p - \mu_p)(X_p - \mu_p)\right] \end{pmatrix} \\
&= \begin{pmatrix} \operatorname{Cov}(X_1, X_1) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_p, X_1) & \cdots & \operatorname{Cov}(X_p, X_p) \end{pmatrix} \\
&= \begin{pmatrix} \operatorname{Var}(X_1) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_p, X_1) & \cdots & \operatorname{Var}(X_p) \end{pmatrix}
\end{aligned}
$$

where:

- the step from the third to fourth line uses the expectation of a random matrix,
- the step from the fourth to fifth line uses Definition 0.1, and
- the last step uses Lemma 0.1.

□

Theorem 0.5 (Alternate expression for variance of a random vector)**.**

$$\operatorname{Var}(\underset{\sim}{X}) = \mathbb{E}\left[\underset{\sim}{X}\underset{\sim}{X}^\top\right] - (\mathbb{E}\underset{\sim}{X})(\mathbb{E}\underset{\sim}{X})^\top$$

Proof

Proof.

$$
\begin{aligned}
\operatorname{Var}(\underset{\sim}{X}) &= \mathbb{E}\left[(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})^\top\right] && \text{(definition)} \\
&= \mathbb{E}\left[\underset{\sim}{X}\underset{\sim}{X}^\top - \underset{\sim}{X}(\mathbb{E}\underset{\sim}{X})^\top - (\mathbb{E}\underset{\sim}{X})\underset{\sim}{X}^\top + (\mathbb{E}\underset{\sim}{X})(\mathbb{E}\underset{\sim}{X})^\top\right] && \text{(expand the product)} \\
&= \mathbb{E}\left[\underset{\sim}{X}\underset{\sim}{X}^\top\right] - (\mathbb{E}\underset{\sim}{X})(\mathbb{E}\underset{\sim}{X})^\top - (\mathbb{E}\underset{\sim}{X})(\mathbb{E}\underset{\sim}{X})^\top + (\mathbb{E}\underset{\sim}{X})(\mathbb{E}\underset{\sim}{X})^\top && \text{(linearity, element-wise; } \mathbb{E}\underset{\sim}{X} \text{ is constant)} \\
&= \mathbb{E}\left[\underset{\sim}{X}\underset{\sim}{X}^\top\right] - (\mathbb{E}\underset{\sim}{X})(\mathbb{E}\underset{\sim}{X})^\top && \text{(combine like terms)}
\end{aligned}
$$

□

Theorem 0.6 (Variance of a linear combination)**.** *For any vector of random variables* $\underset{\sim}{X} = (X_1, \dots, X_n)$ *and corresponding vector of constants* $\underset{\sim}{a} = (a_1, \dots, a_n)$*, the variance of their linear combination is:*

$$\begin{aligned}\operatorname{Var}(\underset{\sim}{a} \cdot \underset{\sim}{X}) &= \operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) \\ &= \underset{\sim}{a}^{\top} \operatorname{Var}(\underset{\sim}{X}) \underset{\sim}{a} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j)\end{aligned}$$

Proof

Proof. Treat $\underset{\sim}{a}$ and $\underset{\sim}{X}$ as $n \times 1$ column vectors, so

$$\begin{aligned}\underset{\sim}{a} \cdot \underset{\sim}{X} &= \underset{\sim}{a}^{\top} \underset{\sim}{X} \\ &= \sum_{i=1}^{n} a_i X_i,\end{aligned}$$

a scalar. By linearity of expectation, $\mathbb{E}\left[\underset{\sim}{a}^{\top} \underset{\sim}{X}\right] = \underset{\sim}{a}^{\top} \mathbb{E}\underset{\sim}{X}$, so:

$$\begin{aligned}
\operatorname{Var}(\underset{\sim}{a}^{\top} \underset{\sim}{X}) &= \mathbb{E}\left[(\underset{\sim}{a}^{\top} \underset{\sim}{X} - \underset{\sim}{a}^{\top} \mathbb{E}\underset{\sim}{X})^2\right] && \text{(definition of variance)} \\
&= \mathbb{E}\left[(\underset{\sim}{a}^{\top} (\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X}))^2\right] && \text{(factor out } \underset{\sim}{a}^{\top}\text{)} \\
&= \mathbb{E}\left[\underset{\sim}{a}^{\top} (\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})^{\top} \underset{\sim}{a}\right] && \text{(a scalar equals its transpose, so } s^2 = s\, s^{\top}\text{)} \\
&= \underset{\sim}{a}^{\top}\, \mathbb{E}\left[(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})(\underset{\sim}{X} - \mathbb{E}\underset{\sim}{X})^{\top}\right] \underset{\sim}{a} && \text{(linearity of expectation, element-wise)} \\
&= \underset{\sim}{a}^{\top} \operatorname{Var}(\underset{\sim}{X}) \underset{\sim}{a} && \text{(variance of a random vector)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j) && \text{(expand the quadratic form, using the elements of } \operatorname{Var}(\underset{\sim}{X})\text{)}
\end{aligned}$$

□

Corollary 0.2 (Variance of a sum of two random variables)**.** *For any two random variables* X *and* Y *and scalars* a *and* b*:*

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2(a \cdot b) \operatorname{Cov}(X, Y)$$

Proof

Proof. Apply Theorem 0.6 with $n = 2$, $X_1 = X$, and $X_2 = Y$:

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

Alternatively, by linearity of expectation:

$$\begin{aligned}
\mathrm{Var}(aX + bY) &\overset{\text{def}}{=} \mathbb{E}[(aX + bY - \mathbb{E}[aX + bY])^2] && \text{(definition of variance)} \\
&= \mathbb{E}[(a(X - \mu_X) + b(Y - \mu_Y))^2] && \text{(linearity of expectation} \\
&= \mathbb{E}[a^2(X - \mu_X)^2 + 2(a \cdot b)(X - \mu_X)(Y - \mu_Y) + b^2(Y - \mu_Y)^2] && \text{(expand the square)} \\
&= a^2\mathbb{E}[(X - \mu_X)^2] + 2(a \cdot b)\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] + b^2\mathbb{E}[(Y - \mu_Y)^2] && \text{(linearity of expectation} \\
&= a^2\mathrm{Var}(X) + 2(a \cdot b)\mathrm{Cov}(X, Y) + b^2\mathrm{Var}(Y) && \text{(definitions of variance a}
\end{aligned}$$

□

Corollary 0.3 (Variance of a sum of independent random variables)**.** *If X and Y are independent, each discrete or continuous, with $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$, then:*

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

Proof

Proof. By Theorem 0.2, $\mathrm{Cov}(X, Y) = 0$. Applying Corollary 0.2 with

$$\begin{aligned}
a &= b \\
&= 1 :
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X + Y) &= 1^2\mathrm{Var}(X) + 1^2\mathrm{Var}(Y) + 2(1 \cdot 1)\mathrm{Cov}(X, Y) && \text{(variance of a sum of two random variables)} \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) && \text{(simplify)} \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) && (\mathrm{Cov}(X, Y) = 0)
\end{aligned}$$

□

Remark

Remark. Corollary 0.2 is why two variables' covariance matters for combining them: the covariance term is what separates the variance of their sum from the sum of their variances, and Corollary 0.3 is the case where that term is 0.

Theorem 0.7 (A variance matrix is symmetric and positive semidefinite)**.** *For a $p \times 1$ random vector $\underset{\sim}{X} = (X_1, \ldots, X_p)^\top$ with $\mathbb{E}[X_i^2] < \infty$ for every i, $\mathrm{Var}(\underset{\sim}{X})$ is symmetric and positive semidefinite: for every $p \times 1$ vector of constants $\underset{\sim}{a}$,*

$$\underset{\sim}{a}^\top \mathrm{Var}(\underset{\sim}{X}) \underset{\sim}{a} \geq 0$$

Proof

Proof. By Theorem 0.4, the (i,j)-th element of $\mathrm{Var}(\underset{\sim}{X})$ is $\mathrm{Cov}(X_i, X_j)$, and by Definition 0.1, with $\mu_i = \mathbb{E}[X_i]$:

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)\big] && \text{(definition of covariance)} \\ &= \mathbb{E}\big[(X_j - \mu_j)(X_i - \mu_i)\big] && \text{(multiplication of numbers is commutative)} \\ &= \mathrm{Cov}(X_j, X_i) && \text{(definition of covariance)} \end{aligned}$$

so the (i,j)-th and (j,i)-th elements are equal, and $\mathrm{Var}(\underset{\sim}{X})$ is symmetric.
For positive semidefiniteness, let $Y = \underset{\sim}{a}^\top \underset{\sim}{X}$. Then:

$$\begin{aligned} \underset{\sim}{a}^\top \mathrm{Var}(\underset{\sim}{X}) \underset{\sim}{a} &= \mathrm{Var}(Y) && \text{(variance of a linear combination)} \\ &= \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big] && \text{(definition of variance)} \\ &\geq 0 && ((Y - \mathbb{E}[Y])^2 \geq 0) \end{aligned}$$

The first step is Theorem 0.6. The last step holds because a random variable that is never negative has a non-negative expectation: in the discrete and continuous cases of the definition of expectation, every term of the sum, or the integrand, is non-negative (for the general case, see Billingsley [4]). □

Example 0.7 (Two variance matrices that differ only in sign)**.** Let $\underset{\sim}{X} = (X_1, X_2)^\top$ be equally likely to be each of the four points

$$D_1 = \{(-5, 1),\ (0, -1),\ (0, 1),\ (5, -1)\},$$

and let $\underset{\sim}{X}'$ be equally likely to be each of the four points

$$D_2 = \{(5, 1),\ (0, -1),\ (0, 1),\ (-5, -1)\}.$$

Both sets have first coordinates $\{-5, 0, 0, 5\}$ and second coordinates $\{1, -1, 1, -1\}$, so

$$\begin{aligned} \mathbb{E}[X_1] &= \mathbb{E}[X_2] \\ &= 0, \end{aligned}$$

$$\begin{aligned}\mathbb{E}[X_1^2] &= (25+0+0+25)/4\\ &= 12.5,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[X_2^2] &= (1+1+1+1)/4\\ &= 1,\end{aligned}$$

and likewise for $\underset{\sim}{X}'$. They differ in the products of the coordinates:

$$\begin{aligned}\mathbb{E}[X_1X_2] &= \frac{(-5)(1)+(0)(-1)+(0)(1)+(5)(-1)}{4}\\ &= -2.5,\\ \mathbb{E}[X_1'X_2'] &= \frac{(5)(1)+(0)(-1)+(0)(1)+(-5)(-1)}{4}\\ &= 2.5.\end{aligned}$$

Since the means are 0, Theorem 0.4 and Theorem 0.1 give:

$$\operatorname{Var}(\underset{\sim}{X}) = \begin{pmatrix}12.5 & -2.5\\ -2.5 & 1\end{pmatrix}, \qquad \operatorname{Var}(\underset{\sim}{X}') = \begin{pmatrix}12.5 & 2.5\\ 2.5 & 1\end{pmatrix}.$$

Both are symmetric. Theorem 0.6 with $\underset{\sim}{a} = (1,1)^\top$ and $\underset{\sim}{a} = (1,-1)^\top$ gives:

$$\begin{aligned}\operatorname{Var}(X_1+X_2) &= 12.5+1+2(-2.5) & \operatorname{Var}(X_1-X_2) &= 12.5+1-2(-2.5)\\ &= 8.5, & &= 18.5,\\ \operatorname{Var}(X_1'+X_2') &= 12.5+1+2(2.5) & \operatorname{Var}(X_1'-X_2') &= 12.5+1-2(2.5)\\ &= 18.5, & &= 8.5.\end{aligned}$$

As a check, X_1+X_2 takes the values $-4,-1,1,4$, each with probability $1/4$, so its mean is 0 and its variance is $(16+1+1+16)/4 = 8.5$.

Remark

Remark. The two sets of points have the same spread along each coordinate axis, so the diagonals of their variance matrices agree. The sign of the covariance says which diagonal direction, $(1,1)^\top$ or $(1,-1)^\top$, has the larger spread.

Example 0.8 (A variance matrix that is not positive definite)**.** Let X_1 have variance $\sigma^2 > 0$, and let $X_2 = X_1$. Every covariance in $\underset{\sim}{X} = (X_1, X_2)^\top$ is $\operatorname{Cov}(X_1, X_1) = \sigma^2$ by Lemma 0.1, so:

$$\operatorname{Var}(\underset{\sim}{X}) = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

With $\underset{\sim}{a} = (1, -1)^\top$,

$$\begin{aligned} \underset{\sim}{a}^\top \operatorname{Var}(\underset{\sim}{X}) \underset{\sim}{a} &= \operatorname{Var}(X_1 - X_2) \\ &= \operatorname{Var}(0) \\ &= 0, \end{aligned}$$

so $\operatorname{Var}(\underset{\sim}{X})$ is positive semidefinite but not positive definite (compare this example). If X_1 is continuous, $\underset{\sim}{X}$ has no joint density.

Corollary 0.4 (Independent components give a diagonal variance matrix)**.** *Let* $\underset{\sim}{X} = (X_1, \dots, X_p)^\top$ *be a random vector whose components are each discrete or continuous, with:*

- $\mathbb{E}[X_i^2] < \infty$ *for every* i*, and*
- X_i *and* X_j *independent for every* $i \neq j$*.*

Then $\operatorname{Var}(\underset{\sim}{X})$ *is diagonal, with diagonal elements* $\operatorname{Var}(X_1), \dots, \operatorname{Var}(X_p)$*.*

Proof

Proof. By Theorem 0.4, the (i, j)-th element of $\operatorname{Var}(\underset{\sim}{X})$ is $\operatorname{Cov}(X_i, X_j)$. For $i \neq j$, X_i and X_j are independent, so $\operatorname{Cov}(X_i, X_j) = 0$ by Theorem 0.2. The (i, i)-th element is $\operatorname{Cov}(X_i, X_i) = \operatorname{Var}(X_i)$ by Lemma 0.1. □

Theorem 0.8 (Correlation lies between -1 and 1)**.** *If* X *and* Y *have finite, positive variances, then their correlation lies in* $[-1, 1]$ *[5]:*

$$-1 \leq \operatorname{Cor}(X, Y) \leq 1$$

Proof

Proof. Write σ_X and σ_Y for the standard deviations,

$$\sigma_X \overset{\text{def}}{=} \operatorname{SD}(X) > 0, \qquad \sigma_Y \overset{\text{def}}{=} \operatorname{SD}(Y) > 0,$$

and take either sign $\pm$ throughout. A variance is the expectation of a squared deviation, a non-negative random variable, so it is non-negative. Applying Corollary 0.2 with $a = 1/\sigma_X$ and $b = \pm 1/\sigma_Y$:

$$
\begin{aligned}
0 &\le \operatorname{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) && \text{(a variance is non-negative)} \\
&= \frac{\operatorname{Var}(X)}{\sigma_X^2} + \frac{\operatorname{Var}(Y)}{\sigma_Y^2} \pm \frac{2\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} && \text{(variance of a sum of two random variables)} \\
&= 1 + 1 \pm \frac{2\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} && \text{(definition of standard deviation: } \sigma_X^2 = \operatorname{Var}(X),\ \sigma_Y^2 = \operatorname{Var}(Y)\text{)} \\
&= 2 \pm 2\operatorname{Cor}(X,Y) && \text{(definition of correlation)}
\end{aligned}
$$

With the $+$ sign, $0 \le 2 + 2\operatorname{Cor}(X,Y)$ gives $\operatorname{Cor}(X,Y) \ge -1$. With the $-$ sign, $0 \le 2 - 2\operatorname{Cor}(X,Y)$ gives $\operatorname{Cor}(X,Y) \le 1$. □

Example 0.9 (Correlation of horsepower and fuel economy)**.** The Auto data record the horsepower and fuel economy (miles per gallon) of 392 car models sold between 1970 and 1982. They come from the StatLib library at Carnegie Mellon University and ship with the book *An Introduction to Statistical Learning*, in the Python package `ISLP` (`ISLP.load_data("Auto")`) and the R package `ISLR2` (`ISLR2::Auto`). This site has a copy at `data/auto.csv`, and both code tabs below read that file.

Choose one car at random, each with the same probability, and let X be its horsepower and Y its miles per gallon. Then each expectation in the definitions of the variance and the covariance is an average over the cars in the data, and Definition 0.2 gives their correlation. The code below computes each part of the definition directly.

Python

```python
import numpy as np
import pandas as pd

auto = pd.read_csv("data/auto.csv")
x = auto["horsepower"].to_numpy(dtype=float)
y = auto["mpg"].to_numpy(dtype=float)

cov_xy = np.mean((x - x.mean()) * (y - y.mean()))
sd_x = np.sqrt(np.mean((x - x.mean()) ** 2))
sd_y = np.sqrt(np.mean((y - y.mean()) ** 2))

pd.Series({
    "Cov(X, Y)": cov_xy,
    "SD(X)": sd_x,
    "SD(Y)": sd_y,
    "Cor(X, Y)": cov_xy / (sd_x * sd_y),
})
#> Cov(X, Y)   -233.261349
#> SD(X)         38.442033
#> SD(Y)          7.795046
#> Cor(X, Y)     -0.778427
#> dtype: float64
```

R

```r
auto <- read.csv("data/auto.csv")
x <- auto$horsepower
y <- auto$mpg

cov_xy <- mean((x - mean(x)) * (y - mean(y)))
sd_x <- sqrt(mean((x - mean(x))^2))
sd_y <- sqrt(mean((y - mean(y))^2))

c(
  "Cov(X, Y)" = cov_xy,
  "SD(X)" = sd_x,
  "SD(Y)" = sd_y,
  "Cor(X, Y)" = cov_xy / (sd_x * sd_y)
)
#>    Cov(X, Y)        SD(X)        SD(Y)    Cor(X, Y)
#> -233.2613494   38.4420327    7.7950458   -0.7784268
```

Figure 3 plots each car's miles per gallon against its horsepower.

```
tibble::tibble(x, y) |>
  ggplot2::ggplot(ggplot2::aes(x = x, y = y)) +
  ggplot2::geom_point(alpha = 0.4, size = 2) +
  ggplot2::labs(x = "horsepower", y = "miles per gallon") +
  ggplot2::theme_minimal()
```

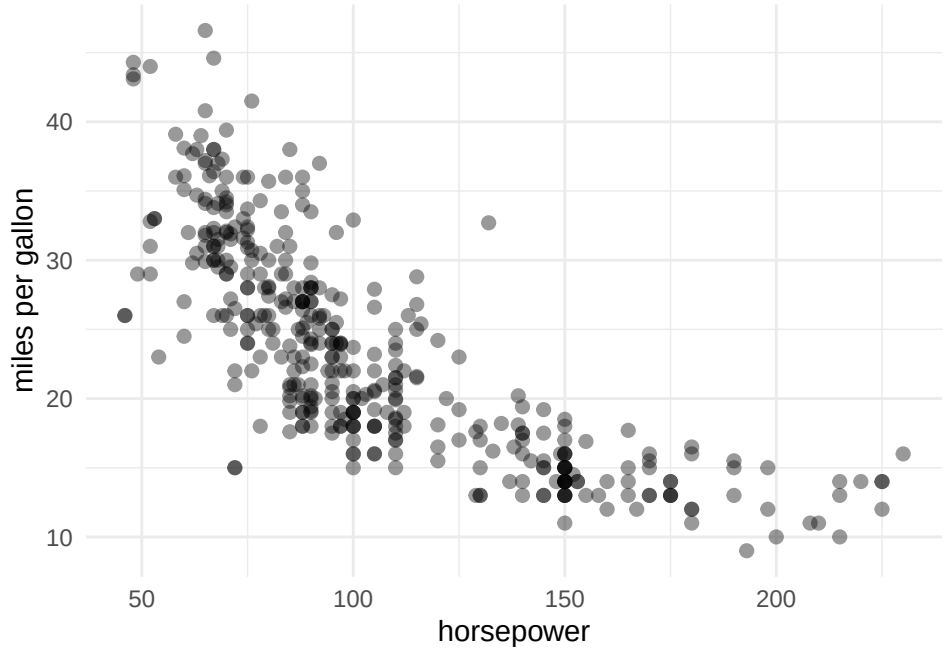


Figure 3: Miles per gallon against horsepower for the cars in the Auto data. Cars with more horsepower tend to have lower fuel economy.

The covariance is -233.3, in units of horsepower times miles per gallon. It is negative, so cars with more horsepower than average tend to have lower fuel economy than average. Dividing by the two standard deviations, 38.4 horsepower and 7.8 miles per gallon, removes the units and gives a correlation of -0.78, which lies between -1 and 1 as Theorem 0.8 requires. The same value comes from R's built-in `cor(x, y)`, because the correlation computed from the data divides the covariance and both standard deviations by the same count, which cancels.

Watch

5MinuteAI explains variance and covariance and correlation in five minutes each, following the University of Washington's CSE 312 [6].

Moments and moment generating functions

Definition 0.1 (Raw moment)**.** The k**-th raw moment** of a random variable X, for a positive integer k with $\mathbb{E}[|X|^k] < \infty$, is the expectation of X^k:

$$\mathbb{E}[X^k]$$

Example 0.1 (Raw moments of a die roll)**.** Let X be the number shown by one roll of a fair six-sided die, so that $\mathbb{P}(X = x) = 1/6$ for each $x \in \{1, 2, 3, 4, 5, 6\}$. The first raw moment is the expectation $\mathbb{E}[X]$, and the second is $\mathbb{E}[X^2]$:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=1}^{6} x \cdot \frac{1}{6} && \text{(definition of expectation)} \\
&= \frac{21}{6} && (1+2+3+4+5+6=21) \\
&= \frac{7}{2} && \text{(simplify)} \\
\mathbb{E}[X^2] &= \sum_{x=1}^{6} x^2 \cdot \frac{1}{6} && \text{(expectation of a function of a discrete random variable)} \\
&= \frac{91}{6} && (1+4+9+16+25+36=91)
\end{aligned}$$

Definition 0.2 (Central moment)**.** The k**-th central moment** of a random variable X, for a positive integer k with $\mathbb{E}[|X|^k] < \infty$, is the expectation of the deviation from the mean raised to the power k:

$$\mathbb{E}[[\mathbb{d}(X)]^k]$$

Example 0.2 (The third central moment of a Bernoulli random variable)**.** Let $X \sim \text{Ber}(\pi)$, so $\mathbb{E}[X] = \pi$ and $\mathbb{d}(X) = X - \pi$. The third central moment is:

$$
\begin{aligned}
\mathbb{E}\big[[\mathrm{d}(X)]^3\big] &= \sum_{x\in\{0,1\}} (x-\pi)^3 \cdot \mathbb{P}(X=x) && \text{(expectation of a function of a discrete random variable)}\\
&= (0-\pi)^3(1-\pi) + (1-\pi)^3\pi && \text{(substitute the Bernoulli probabilities)}\\
&= \pi(1-\pi)((1-\pi)^2 - \pi^2) && \text{(factor out } \pi(1-\pi)\text{)}\\
&= \pi(1-\pi)(1-2\pi) && \text{(difference of squares: } (1-\pi)^2-\pi^2 = 1-2\pi\text{)}
\end{aligned}
$$

For a fair coin ($\pi = 1/2$), the third central moment is 0.

Theorem 0.1 (The second central moment is the variance)**.** *If* $\mathbb{E}[X^2] < \infty$*, then the second central moment of* X *is* $\mathrm{Var}(X)$*.*

Proof

Proof. The second central moment is $\mathbb{E}\big[[\mathrm{d}(X)]^2\big]$ by Definition 0.2, which is $\mathrm{Var}(X)$ by Definition 0.1. □

Example 0.3 (The variance of a die roll as a central moment)**.** For the die roll X of Example 0.1, $\mathbb{E}[X] = 7/2$ and $\mathbb{E}[X^2] = 91/6$. By Theorem 0.1, the second central moment of X is the variance $\mathrm{Var}(X)$, which Theorem 0.2 gives as:

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && \text{(simplified expression for variance)}\\
&= \frac{91}{6} - \left(\frac{7}{2}\right)^2 && \text{(substitute the raw moments)}\\
&= \frac{182}{12} - \frac{147}{12} && \text{(common denominator)}\\
&= \frac{35}{12} && \text{(subtract)}
\end{aligned}
$$

Definition 0.3 (Moment generating function)**.** The **moment generating function** of a random variable X is the function M_X of a real number t that gives the expectation of e^{tX}, defined at every t with $\mathbb{E}[\mathrm{e}^{tX}] < \infty$:

$$M_X(t) \stackrel{\text{def}}{=} \mathbb{E}[\mathrm{e}^{tX}]$$

Example 0.4 (Moment generating function of a Bernoulli random variable)**.** Let $X \sim \mathrm{Ber}(\pi)$, so $\mathbb{P}(X=1) = \pi$ and $\mathbb{P}(X=0) = 1-\pi$. For every t:

$$
\begin{aligned}
M_X(t) &= \mathbb{E}[\mathrm{e}^{tX}] && \text{(definition of the moment generating function)} \\
&= \sum_{x \in \{0,1\}} \mathrm{e}^{tx} \cdot \mathbb{P}(X = x) && \text{(expectation of a function of a discrete random variable)} \\
&= \mathrm{e}^0 \cdot (1 - \pi) + \mathrm{e}^t \cdot \pi && \text{(substitute the Bernoulli probabilities)} \\
&= 1 - \pi + \pi\,\mathrm{e}^t && (\mathrm{e}^0 = 1)
\end{aligned}
$$

Theorem 0.2 (Moments from the moment generating function)**.** *If $M_X(t)$ is defined at every t in some open interval around 0, then for every positive integer k, the k-th raw moment of X is the k-th derivative of M_X at 0:*

$$\mathbb{E}[X^k] = M_X^{(k)}(0)$$

Proof

Proof. Let $h > 0$ be such that $M_X(t)$ is defined for every $|t| < h$. For any such t, the series of the exponential gives:

$$\mathrm{e}^{tX} = \sum_{k=0}^{\infty} \frac{t^k X^k}{k!}$$

The absolute value of each summand is at most the matching summand of $\mathrm{e}^{|tX|}$. Also, $\mathrm{e}^{|tX|} \le \mathrm{e}^{tX} + \mathrm{e}^{-tX}$, and $M_X(t)$ and $M_X(-t)$ are both defined, so $\mathbb{E}\left[\mathrm{e}^{|tX|}\right] < \infty$. This domination lets the expectation pass through the series, by the Fubini–Tonelli exchange:

$$
\begin{aligned}
M_X(t) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{t^k X^k}{k!}\right] && \text{(power series of the exponential)} \\
&= \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}[X^k] && \text{(pass the expectation through the sum; linearity of expectation)}
\end{aligned}
$$

A Taylor series can be differentiated one summand at a time inside the open interval where it converges. The k-th derivative at 0 keeps only the summand with t^k, and that summand's coefficient is multiplied by $k!$. So $M_X^{(k)}(0) = k! \cdot \frac{\mathbb{E}[X^k]}{k!} = \mathbb{E}[X^k]$. □

Example 0.5 (Moments of a Bernoulli random variable from its moment generating function)**.** For $X \sim \mathrm{Ber}(\pi)$, Example 0.4 gives $M_X(t) = 1 - \pi + \pi\,\mathrm{e}^t$ for every t, so $M_X^{(k)}(t) = \pi\,\mathrm{e}^t$ for every $k \ge 1$. By Theorem 0.2, for every $k \ge 1$:

$$\begin{aligned}
\mathbb{E}[X^k] &= M_X^{(k)}(0) && \text{(moments from the moment generating function)} \\
&= \pi\, \mathrm{e}^0 && (M_X^{(k)}(t) = \pi\, \mathrm{e}^t) \\
&= \pi && (\mathrm{e}^0 = 1)
\end{aligned}$$

This agrees with the direct computation of $\mathbb{E}[X^2]$ for a Bernoulli variable on the expectation page.

Theorem 0.3 (Moment generating function of a sum of independent random variables)**.** *If X and Y are independent, and $M_X(t)$ and $M_Y(t)$ are defined at a number t, then:*

$$M_{X+Y}(t) = M_X(t)\, M_Y(t)$$

Proof

Proof. Functions of independent variables are independent, so e^{tX} and e^{tY} are independent. Independent variables A and B satisfy $\mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B]$, so:

$$\begin{aligned}
M_{X+Y}(t) &= \mathbb{E}\left[\mathrm{e}^{t(X+Y)}\right] && \text{(definition of the moment generating function)} \\
&= \mathbb{E}\left[\mathrm{e}^{tX}\, \mathrm{e}^{tY}\right] && (\mathrm{e}^{a+b} = \mathrm{e}^a\, \mathrm{e}^b) \\
&= \mathbb{E}\left[\mathrm{e}^{tX}\right] \mathbb{E}\left[\mathrm{e}^{tY}\right] && \text{(expectation of a product of independent variables)} \\
&= M_X(t)\, M_Y(t) && \text{(definition of the moment generating function)}
\end{aligned}$$

□

Example 0.6 (Moment generating function of a binomial random variable)**.** By the sum of Bernoulli variables, a binomial random variable $S \sim \mathrm{Bin}(n, \pi)$ is a sum $S = \sum_{i=1}^n X_i$ of n independent $\mathrm{Ber}(\pi)$ variables. Each has moment generating function $1 - \pi + \pi\, \mathrm{e}^t$ (Example 0.4). Applying Theorem 0.3 one summand at a time:

$$\begin{aligned}
M_S(t) &= \prod_{i=1}^n M_{X_i}(t) && \text{(moment generating function of a sum of independent random variables)} \\
&= (1 - \pi + \pi\, \mathrm{e}^t)^n && \text{(each factor is the Bernoulli moment generating function)}
\end{aligned}$$

References

[1] Wikipedia contributors, "Errors and residuals — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Errors_and_residuals

[2] Wikipedia contributors, "Variance — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Variance

[3] Wikipedia contributors, "Covariance — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Covariance

[4] P. Billingsley, *Probability and measure*, 3rd ed. in Wiley series in probability and mathematical statistics. New York: Wiley, 1995.

[5] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[6] 5MinuteAI, "Probability & statistics." Video playlist on YouTube. Accessed: Oct. 12, 2026. [Online]. Available: https://www.youtube.com/playlist?list=PLeB45KifGiuHesi4PALNZSYZFhViVGQJK