CHAPTER 2

# Describing Political Data

GOVT 10: Quantitative Political Analysis

This chapter introduces the fundamental tools for summarizing and understanding political data. We begin with measures of central tendency, exploring when the mean provides useful information and when the median offers a more honest picture of ”typical” values. We then turn to measures of spread, examining why variance and standard deviation matter for understanding how diverse a population or sample really is. The 68-95-99.7 rule provides a powerful tool for interpreting normal distributions, while correlation allows us to quantify relationships between variables. Throughout, we emphasize a critical theme that will recur across this course: observing that two variables move together tells us nothing about whether one causes the other.

## 1 The Question of “Typical”

Political debates frequently hinge on how we characterize what is normal, average, or typical. When a labor union argues that wages have stagnated, they point to one set of numbers. When a business association counters that compensation has grown, they point to another. Both may be drawing from the same underlying data. The difference lies in how they summarize it.

Consider the question of wealth in a democracy. How wealthy is the “typical” citizen? The answer has profound implications for tax policy, social programs, and our understanding of economic inequality. If we want to know whether the average person is doing well, we need to be precise about what “average” means.

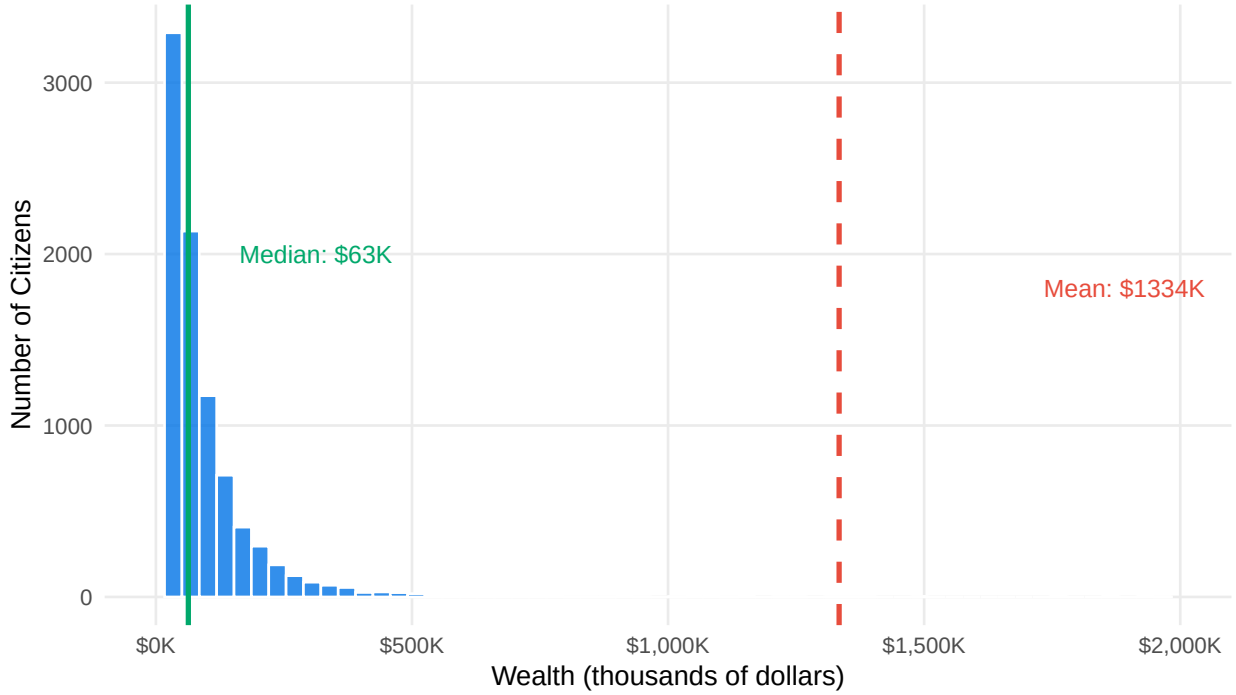


Figure 1: Figure 2.1: Wealth distribution in a hypothetical democracy. The long right tail from a small number of very wealthy individuals pulls the mean far above where most citizens actually fall.

Figure 2.1 illustrates this challenge. In this hypothetical democracy, a small number of extremely wealthy individuals pull the arithmetic mean far above where most citizens actually fall. The mean wealth is over $1,334 thousand, but the median citizen has only about $63 thousand. If a politician claimed that "the average citizen has substantial wealth," they would be technically correct but deeply misleading. The median tells a more honest story about the economic reality facing ordinary people.

This distinction between mean and median is not merely academic. It shapes how we think about economic policy, whether inequality is increasing or decreasing, and who benefits from different political arrangements. Throughout this chapter, we will develop the tools to describe data accurately and honestly.

## 2 Measures of Central Tendency

When we describe a dataset, we typically want to identify some "central" or "representative" value. The three most common measures are the mean, the median, and the mode. Each captures a different sense of what is typical, and choosing among them requires understanding the shape of our data.

## 2.1 The Arithmetic Mean

The mean is calculated by summing all observations and dividing by their count:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

We can think of the mean as the "balance point" of a distribution. If we placed each observation on a seesaw according to its value, the mean is where the seesaw would balance. This physical intuition helps explain why the mean is sensitive to extreme values: a very heavy weight far from the center exerts strong leverage, pulling the balance point toward it.

```
legislatures <- c(40, 49, 50, 60, 67, 80, 99, 100, 120, 400)
mean(legislatures)
```

```
[1] 106.5
```

> **What this code does:** We create a vector containing the number of seats in ten state legislatures. The `mean()` function adds all values together and divides by the count. The result of 106.5 tells us the average legislature size, though one large value (400) heavily influences this number.

The mean of 106.5 is heavily influenced by one very large legislature with 400 seats. This single observation pulls the mean well above where most legislatures actually fall.

## 2.2 The Median

The median is the value that divides an ordered dataset exactly in half. Fifty percent of observations lie below the median, and fifty percent lie above. To find it, we first sort the data from smallest to largest. If we have an odd number of observations, the median is the middle value. If we have an even number, we average the two middle values.

```
sort(legislatures)
```

```
 [1]  40  49  50  60  67  80  99 100 120 400
```

```
median(legislatures)
```

```
[1] 73.5
```

> **What this code does:** The `sort()` function arranges values in ascending order so we can see where the middle falls. The `median()` function identifies the central value. With ten observations, the median is the average of the 5th and 6th values: (67 + 80) / 2 = 73.5 seats.

The median of 73.5 is much lower than the mean of 106.5. This happens because the median is resistant to extreme values. Whether the largest legislature has 400 seats or 4,000 seats, the median remains unchanged. This resistance makes the median particularly useful when we have outliers or skewed distributions.

### 2.3 Choosing Between Mean and Median

The choice between mean and median depends on what question we are trying to answer and the shape of our data. When a distribution is roughly symmetric and free of extreme outliers, the mean and median will be close to each other, and either provides a reasonable summary. When a distribution is skewed or contains outliers, they diverge, and we must think carefully about which is more appropriate.

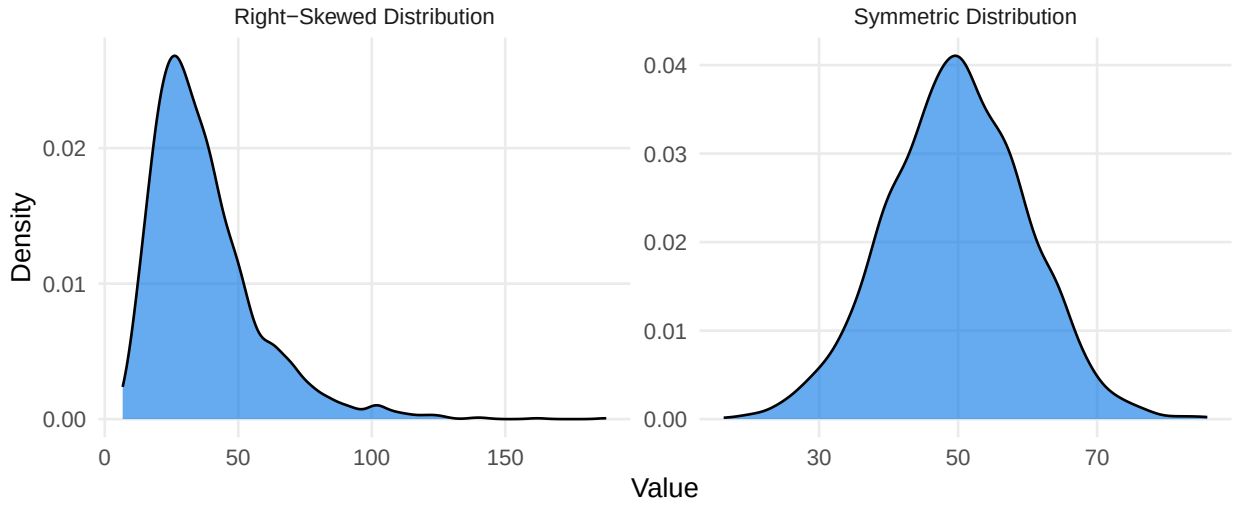


Figure 2: Figure 2.2: Comparing symmetric and skewed distributions. In symmetric distributions, mean and median coincide. In skewed distributions, they diverge.

The table below provides guidance for when to use each measure.

| **Use the Mean When...** | **Use the Median When...** |
|---|---|
| Distribution is roughly symmetric | Distribution is skewed |
| No extreme outliers present | Outliers may distort the picture |
| Mathematical properties are needed | ”Typical” value is the goal |
| Preparing for statistical inference | Communicating to general audiences |

In political science, many variables are skewed: campaign spending, population of electoral districts, wealth, and news attention all have long right tails. For these variables, the median typically provides a more honest characterization of typical values.

## 3 Measures of Spread

Knowing the center of a distribution is only part of the story. Two datasets can have identical means but look completely different. To capture how dispersed or concentrated our observations are, we need measures of spread.

## 3.1 The Range

The simplest measure of spread is the range: the difference between the largest and smallest values. The range is easy to calculate and interpret, but it has a significant weakness: it depends entirely on the two most extreme observations, ignoring everything in between.

```
turnout <- c(52, 55, 58, 59, 61, 62, 63, 64, 65, 67, 68, 71)
range(turnout)
```

```
[1] 52 71
```

```
max(turnout) - min(turnout)
```

```
[1] 19
```

> **What this code does:** We create turnout percentages for 12 precincts. The `range()` function returns both the minimum and maximum. Subtracting min from max gives us 19 percentage points of spread. However, this measure would change dramatically if we added one unusual precinct.

## 3.2 Variance and Standard Deviation

To measure spread in a way that uses all our data, we turn to variance and standard deviation. The core idea is to measure how far, on average, each observation falls from the mean.

We might first think to simply average the distances from the mean. The problem is that distances above the mean are positive while distances below are negative, and they cancel out. By definition, the sum of deviations from the mean equals zero.

```
mean_turnout <- mean(turnout)
distances <- turnout - mean_turnout
sum(distances)
```

```
[1] -2.842171e-14
```

> **What this code does:** We subtract the mean from each observation to get deviations. When we sum these deviations, positive and negative values cancel out, giving us zero (or a tiny rounding error). This demonstrates why simple averaging of distances does not work as a measure of spread.

To solve this problem, we square each distance before averaging. Squaring makes all values positive and also gives extra weight to observations far from the mean, which is often desirable. The result is the variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2}$$

```
squared_distances <- distances^2
n <- length(turnout)
variance <- sum(squared_distances) / (n - 1)
variance
```

```
[1] 30.08333
```

```
var(turnout)
```

```
[1] 30.08333
```

> **What this code does:** We square each deviation to make all values positive, then sum them and divide by $n - 1$ (Bessel's correction). This gives the variance. The `var()` function performs this calculation automatically. The result is in squared units.

The variance is in squared units, which can be hard to interpret. If turnout is measured in percentages, the variance is in "squared percentages," which has no intuitive meaning. The standard deviation solves this problem by taking the square root of variance, returning us to the original units:

$$s = \sqrt{s^2} \tag{3}$$

```
sd(turnout)
```

```
[1] 5.484828
```

```
sqrt(var(turnout))
```

```
[1] 5.484828
```

> **What this code does:** The `sd()` function calculates standard deviation directly. We verify it equals the square root of variance. The result of about 5.4 percentage points tells us that typical turnout rates deviate from the mean by roughly 5 to 6 percentage points.

## 4 The Normal Distribution and the 68-95-99.7 Rule

Many phenomena in nature and society approximate a particular mathematical shape called the normal distribution, also known as the bell curve. The normal distribution is symmetric, with observations clustering around the mean and becoming progressively rarer as we move toward the extremes.

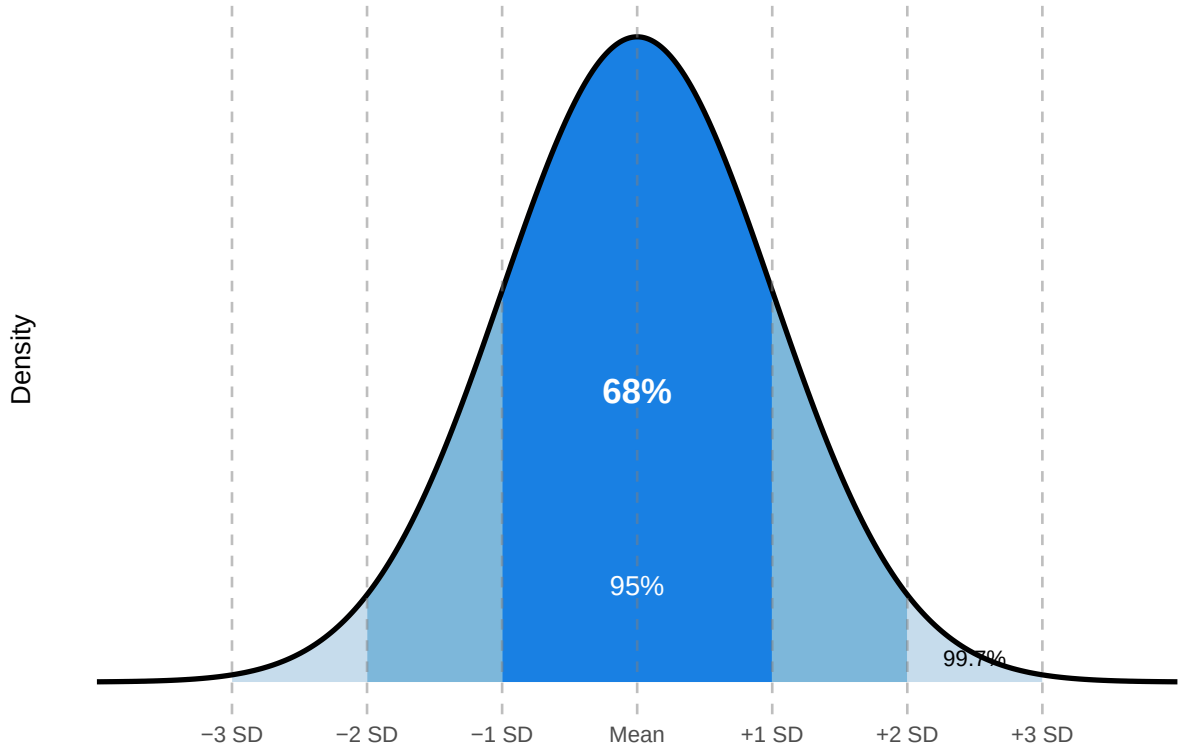


Figure 3: Figure 2.3: The 68-95-99.7 rule. For any normal distribution, approximately 68% of observations fall within one standard deviation of the mean, 95% within two, and 99.7% within three.

For any normal distribution, regardless of its mean or standard deviation, a powerful rule applies. Approximately 68% of observations fall within one standard deviation of the mean. Approximately 95% fall within two standard deviations. And approximately 99.7% fall within three standard deviations. This is the 68-95-99.7 rule, and it provides a quick way to assess whether an observation is unusual.

Suppose we know that support for a policy has a mean of 55% with a standard deviation of 8 percentage points, and that this support is normally distributed across states. We can immediately say that about 68% of states have support between 47% and 63% (one standard deviation on either side). About 95% of states fall between 39% and 71%. A state with 80% support would be more than three standard deviations above the mean, making it exceptionally unusual.

```
mean_support <- 55
sd_support <- 8
c(mean_support - sd_support, mean_support + sd_support)
```

```
[1] 47 63
```

```
c(mean_support - 2*sd_support, mean_support + 2*sd_support)
```

```
[1] 39 71
```

**Worked Example: Is This Observation Unusual?**

**Question:** Voter turnout in a state was 78%. National mean turnout is 60% with SD = 7%. Is this state's turnout unusual?

**Step 1: Calculate how many SDs from the mean.**

$$z = \frac{\text{Observation} - \text{Mean}}{\text{SD}} = \frac{78 - 60}{7} = 2.57$$

**Step 2: Apply the 68-95-99.7 rule.**

- Within 1 SD (53–67%): Normal, 68% of states here
- Within 2 SDs (46–74%): Still common, 95% of states here
- Within 3 SDs (39–81%): Nearly all, 99.7% here

**Step 3: Interpret.** 78% is between 2 and 3 SDs above the mean. This is in the top 2.5% of states.

**Conclusion:** "This state's 78% turnout is 2.57 standard deviations above the national mean—an unusually high turnout that would be expected in fewer than 3% of states."

**Quick rule:** Beyond 2 SDs = unusual. Beyond 3 SDs = rare.

**What this code does:** We define mean and standard deviation for policy support. Subtracting and adding one SD gives the 68% interval (47 to 63). Subtracting and adding two SDs gives the 95% interval (39 to 71). Values outside these ranges are increasingly unusual.

## 5 Correlation

So far we have focused on summarizing single variables. But much of political science concerns relationships between variables. Does media coverage affect candidate support? Is economic growth associated with incumbent success? Does education predict political participation? To address these questions, we need a way to measure how two variables move together.

Correlation measures the strength and direction of a linear relationship between two variables. The correlation coefficient, denoted $r$, ranges from -1 to +1. A positive correlation means that as one variable increases, the other tends to increase as well. A negative correlation means that as one variable increases, the other tends to decrease. A correlation near zero indicates no linear relationship.

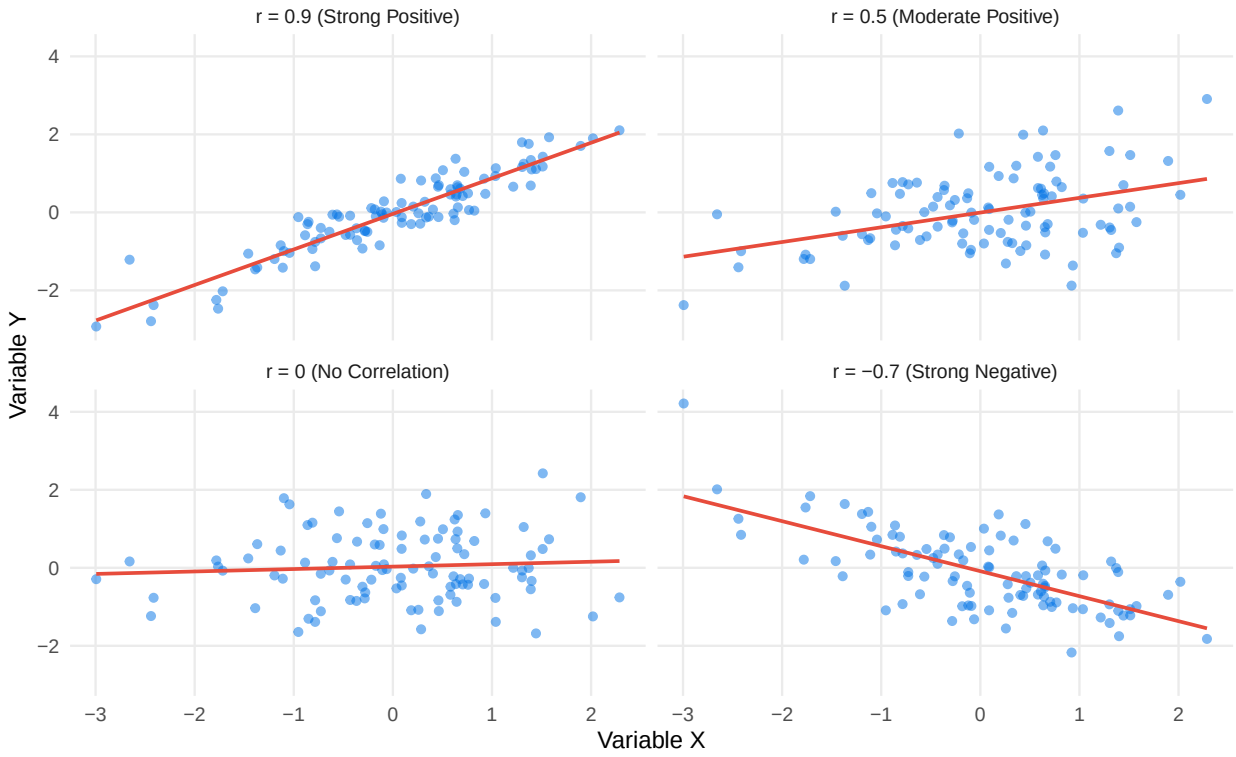


Figure 4: Figure 2.4: Scatterplots showing different correlation strengths. Strong positive correlations show points clustering tightly around an upward slope; strong negative correlations cluster around a downward slope; zero correlation shows no pattern.

The formula for Pearson's correlation coefficient captures the degree to which X and Y vary together, standardized by how much each varies individually:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

In practice, we calculate correlation using the `cor()` function rather than this formula.

```
cor(states_policy$urbanization, states_policy$policy_score)
```

```
[1] 0.7829285
```

**Interpreting Correlation Output Step-by-Step**

**Step 1: Check the sign.**

- Positive (e.g., +0.65): As X increases, Y tends to increase
- Negative (e.g., -0.45): As X increases, Y tends to decrease

**Step 2: Assess the magnitude.**

- $|r| < 0.3$: Weak relationship
- $0.3 \leq |r| < 0.7$: Moderate relationship
- $|r| \geq 0.7$: Strong relationship

**Step 3: Remember the limits.**

- Correlation does NOT imply causation
- Correlation only measures LINEAR relationships
- Outliers can dramatically inflate or deflate $r$

**Template:** "There is a [weak/moderate/strong] [positive/negative] relationship between X and Y (r = [value]). However, this correlation does not establish that X causes Y."

## 5.1 What Correlation Does Not Tell Us

Correlation has critical limitations that every political scientist must understand. Most importantly, correlation does not imply causation. Observing that two variables move together tells us nothing about whether one causes the other, whether both are caused by something else, or whether the relationship is coincidental.

Consider the relationship between per-capita chocolate consumption and the number of Nobel Prize winners across countries. There is a surprisingly strong positive correlation. Does eating chocolate make populations smarter? Almost certainly not. Both variables are probably driven by national wealth, educational investment, and research infrastructure. The correlation is real, but the causal story is spurious.

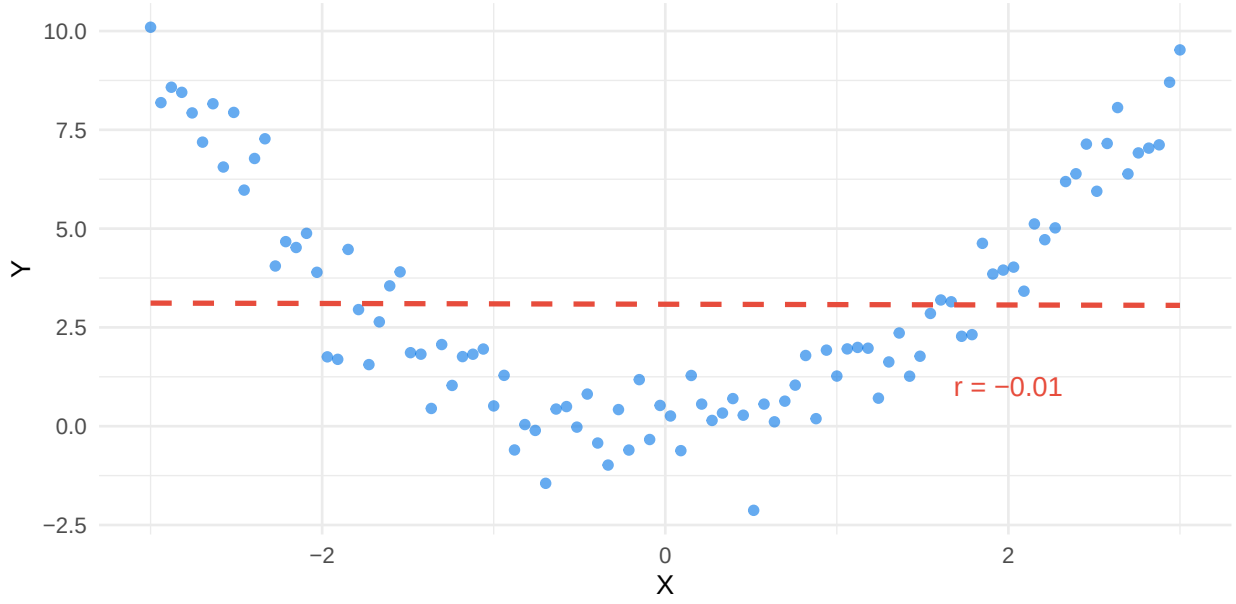


Figure 5: Figure 2.5: A nonlinear relationship with zero correlation. The U-shaped pattern is completely missed by correlation, which only measures linear relationships.

Additionally, correlation only measures linear relationships. Figure 2.5 shows a clear U-shaped relationship between two variables, but the correlation is essentially zero. The variables are strongly related, but not in a straight-line way. Always visualize your data before relying on correlation alone.

> **Critical Misconception: Correlation Implies Causation**
>
> This error appears constantly in media reports and casual political discourse. When you see a claim that X is "linked to" or "associated with" Y, remember that association alone cannot establish causation. At least three possibilities exist: X causes Y, Y causes X, or some third variable Z causes both. Only carefully designed research, particularly randomized experiments, can distinguish among these explanations. We will return to this theme throughout the course.

# 6 Working with Data in R

The concepts we have discussed become practical through the tidyverse, a collection of R packages designed for data analysis. The tidyverse emphasizes readable code that can be understood as a series of steps, much like instructions to a research assistant.

## 6.1 The Pipe Operator

The pipe operator `%>%` passes the result of one operation to the next. This allows us to chain together multiple steps in a way that reads naturally from left to right.

```
round(mean(c(55, 58, 62, 49, 71)), 1)
```

```
[1] 59
```

```
c(55, 58, 62, 49, 71) %>% mean() %>% round(1)
```

```
[1] 59
```

> **What this code does:** The first version nests functions, requiring us to read from inside out. The piped version reads like a sentence: "Take these values, then calculate the mean, then round to one decimal place." Both produce 59, but the pipe is easier to read and modify.

## 6.2 Filtering and Selecting

The `filter()` function keeps rows that meet specified conditions, while `select()` chooses which columns to retain.

```
municipalities %>%
  filter(population > 100000, mayor_party == "Progressive") %>%
  select(municipality, population, mayor_party) %>%
  head(5)
```

```
# A tibble: 5 x 3
  municipality population mayor_party
  <chr>             <dbl> <chr>
1 City_9           409261 Progressive
2 City_20          177041 Progressive
3 City_25          353003 Progressive
4 City_33          125759 Progressive
5 City_48          205443 Progressive
```

> **What this code does:** We start with the municipalities data, then `filter()` keeps only rows where population exceeds 100,000 AND mayor is Progressive. Then `select()` keeps only three columns. Finally, `head(5)` shows only the first five rows.

### 6.3 Grouping and Summarizing

The true power of the tidyverse emerges when we calculate statistics by group. The `group_by()` function divides data into groups, and `summarise()` calculates summary statistics within each group.

```
municipalities %>%
  group_by(region) %>%
  summarise(
    n_cities = n(),
    avg_unemploy = mean(unemployment),
    med_income = median(median_income),
    .groups = "drop"
  )
```

```
# A tibble: 4 x 4
  region n_cities avg_unemploy med_income
  <chr>     <int>        <dbl>      <dbl>
1 East         49         5.91     54508
2 North        51         6.26     51370
3 South        62         5.26     50246.
4 West         38         6.2      49390.
```

> **What this code does:** We group municipalities by region, then calculate three statistics within each group: count (`n()`), mean unemployment, and median income. The `.groups = "drop"` removes the grouping structure afterward, producing a clean summary table.

These tools allow us to quickly answer substantive questions. Which regions have the highest unemployment? Where are incomes concentrated? How do different types of mayors compare? By combining filtering, grouping, and summarizing, we can explore our data efficiently and thoroughly.

## 7 Common Mistakes and How to Avoid Them

> **Mistake 1: Using the Mean for Skewed Distributions**
>
> When data are heavily skewed, the mean can be deeply misleading. Before reporting any average, visualize the distribution. If you see a long tail in one direction, the median is

almost always more appropriate for describing what is "typical." This is especially important for variables like income, wealth, and campaign spending.

### Mistake 2: Ignoring Outliers

A single extreme observation can dramatically affect the mean, standard deviation, and correlation coefficient. Always examine your data for unusual values. An unexpected outlier might indicate a data entry error, or it might be a genuine but unusual case that deserves separate attention. Either way, you should understand its influence on your results.

### Mistake 3: Treating Correlation as Causation

This cannot be emphasized enough. When you observe that two variables are correlated, you have learned something about patterns in your data, but nothing about causation. The correlation between television ownership and life expectancy across countries is high, but televisions do not extend lifespans. Both reflect national development. Distinguishing correlation from causation requires research design, which we will study in Chapters 3 and 4.

### Mistake 4: Forgetting About Missing Data

Missing values in R appear as `NA`. If you calculate the mean of a vector containing `NA`, R will return `NA` rather than ignoring the missing value. Always use `na.rm = TRUE` when appropriate, and always report how much data is missing. A mean calculated from 90% of your observations may be fine; a mean calculated from 10% is probably not representative.

# 8 R Functions Reference

| Function | Purpose | Example |
|---|---|---|
| `mean(x)` | Calculate arithmetic mean | `mean(turnout, na.rm = TRUE)` |
| `median(x)` | Calculate median | `median(income)` |
| `var(x)` | Calculate variance | `var(scores)` |
| `sd(x)` | Calculate standard deviation | `sd(approval)` |
| `range(x)` | Find minimum and maximum | `range(ages)` |
| `cor(x, y)` | Calculate correlation | `cor(spending, votes)` |
| `sum(x)` | Sum all values | `sum(votes)` |
| `length(x)` | Count observations | `length(respondents)` |

Table 1: Summary Statistics Functions

| Function | Purpose | Example |
|---|---|---|
| `filter()` | Keep rows matching conditions | `filter(party == "D")` |
| `select()` | Keep specified columns | `select(name, vote)` |
| `arrange()` | Sort rows | `arrange(desc(votes))` |
| `mutate()` | Create new columns | `mutate(margin = dem - rep)` |
| `group_by()` | Group data for calculations | `group_by(state)` |
| `summarise()` | Calculate summary statistics | `summarise(avg = mean(x))` |

Table 2: Tidyverse Data Manipulation Functions

# 9 Practice Problems

The following questions are designed to test your understanding of the concepts in this chapter. Work through them carefully, and check your reasoning against the principles we have discussed.

**Question 1**

A news report claims that ”the average net worth of members of Congress is $7.2 million.” What measure of central tendency is this likely referring to? Would you expect the median net worth to be higher or lower? Why might the median be more informative for understanding how wealthy the ”typical” member of Congress is?

**Question 2**

Two polling organizations both report that the president’s approval rating is 48%. However, Poll A was conducted in a single state while Poll B was conducted nationally. Which poll

would you expect to have a larger standard deviation in its individual responses? Explain your reasoning in terms of population heterogeneity.

### Question 3

Researchers find a correlation of $r = 0.72$ between a country's per-capita coffee consumption and its scores on international math tests. Propose three different explanations for this relationship that do not involve coffee directly improving mathematical ability.

### Question 4 (Computational)

Using R, create a vector of 15 values representing hypothetical approval ratings: `c(42, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 57, 60, 85, 88)`. Calculate the mean and median. Then remove the last two values (85 and 88) and recalculate. Which measure changed more dramatically, and why?

### For Further Study

This chapter has introduced the foundational tools for describing data. Students interested in going deeper might explore the concept of robust statistics, which provides measures of center and spread that are less sensitive to outliers than the mean and standard deviation. The interquartile range (IQR), for example, measures spread using only the middle 50% of data, ignoring extremes entirely. For more on this topic, consider Wilcox's *Introduction to Robust Estimation and Hypothesis Testing*. For an accessible treatment of statistical thinking more broadly, Wheelan's *Naked Statistics* provides an engaging overview without heavy mathematics.