Week 1: Populations, Samples, and Descriptive Statistics

Introductory Statistics Notes

1 A First Look at Data

Statistics is about using data from a *sample* to learn about a larger *population*. In this course, we will get used to:

- talking about *populations* and *samples*,
- using a few standard symbols ($\mu, \sigma, p, \rho, \bar{x}, s, \hat{p}, r$),
- and describing data with graphs and summary numbers.

Running example: study hours and quiz scores

We will use one main example throughout this week.

Imagine a large community college with many students taking the same introductory statistics course. For each student we record:

- X = number of hours they studied for Quiz 1,
- Y = their score on Quiz 1 (out of 10).

We cannot easily measure every student, so we collect a simple random sample of $n = 6$ students and record their hours and scores:

Student i	1	2	3	4	5	6
Study hours x_i	1	2	2	4	5	6
Quiz score y_i	4	5	6	7	8	9

This small table is our **sample data**.

2 Populations, Samples, and Notation

2.1 Populations

A **population** is the entire group we care about. Examples:

- all students at the college this term,
- all parts coming off a manufacturing line today,
- all current customers of a company.

Numbers that describe a population are called **parameters**. Standard symbols:

- μ (mu): population mean,
- σ (sigma): population standard deviation,
- p: population proportion,

- ρ (rho): population correlation between two variables.

These are usually *unknown*. We will try to estimate them from data.

2.2 Samples

A **sample** is the part of the population that we actually observe.

Numbers computed from a sample are called **statistics**. Standard symbols:

- $\bar{x}$: sample mean,
- s: sample standard deviation,
- $\hat{p}$: sample proportion,
- r: sample correlation.

Each statistic is used to estimate a parameter:

Population (unknown)	Sample (computed)	Meaning
μ	$\bar{x}$	average (mean)
σ	s	spread (standard deviation)
p	$\hat{p}$	proportion
ρ	r	linear association (correlation)

Running example: population vs sample

- **Population:** all students taking intro statistics this term.
- **Sample:** the 6 students in our table.
- **Parameters of interest:** the population mean quiz score μ, and possibly the population correlation ρ between study hours and quiz scores.
- **Statistics:** the sample mean quiz score $\bar{y}$, sample standard deviation s_y, and sample correlation r, computed from the 6 students.

3 Displaying Data: Tables, Bar Charts, Histograms, and Scatterplots

Before we calculate formulas, we want to *see* the data.

3.1 Frequency table and bar chart (one variable)

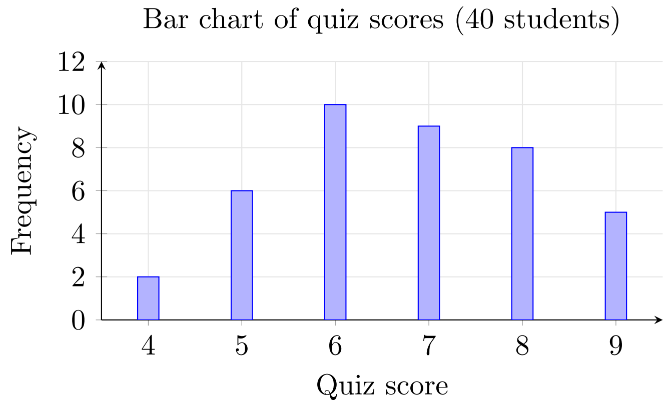
First, look at the distribution of quiz scores alone, ignoring study hours.

Our original sample had 6 students, but imagine we have now collected quiz scores from 40 students. Suppose the scores from 4 to 9 occur with the following frequencies:

Quiz score	4	5	6	7	8	9
Frequency	2	6	10	9	8	5

This is a **frequency table** for a single variable (quiz score). It already tells us something about the shape of the distribution: scores near 6–8 are most common, and low scores (4) and high scores (9) are less common.

We can turn this into a bar chart.



Each bar shows how many students had that exact score. We can quickly see where the distribution peaks and how spread out it is.

3.2 Histogram (grouping values into wider bins)

If we had many possible scores and many students, listing every single score might be too detailed. Instead, we can group nearby scores into **bins** and draw a **histogram**.

For quiz scores on a 0–10 scale, we might choose wider bins that combine scores:

$$4\text{–}5, \quad 6\text{–}7, \quad 8\text{–}9.$$

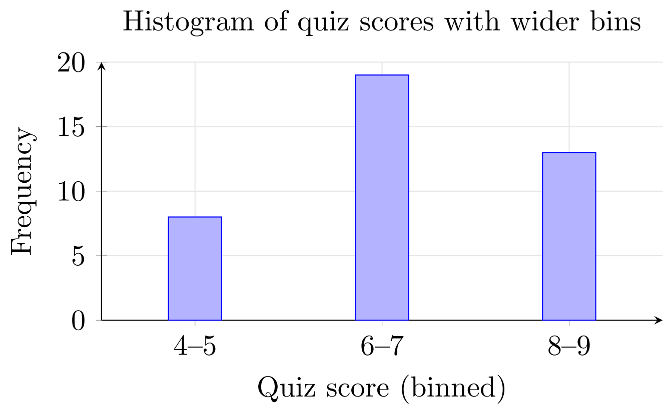
Using the frequencies above:

- Bin 4–5: $2 + 6 = 8$ students,
- Bin 6–7: $10 + 9 = 19$ students,
- Bin 8–9: $8 + 5 = 13$ students.

We can summarize this in a small bin-frequency table:

Bin	Scores included	Frequency
4–5	4 and 5	8
6–7	6 and 7	19
8–9	8 and 9	13

Now the histogram uses these wider bins:



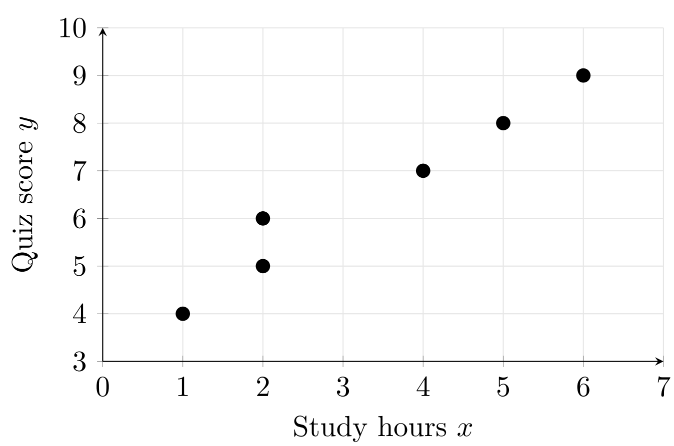
This histogram looks different from the bar chart, even though it is based on the *same* data. By choosing wider bins, we see a smoother, more "summarized" shape: most scores are in the middle (6–7), with fewer in the lower (4–5) and higher (8–9) ranges.

As we move further in the course, histograms will help us see:

- where the center is,
- how spread out the data are,
- whether the distribution is symmetric, skewed, or has multiple peaks,
- whether there are outliers.

3.3 Scatterplot (two variables together)

To see the relationship between *study hours* and *quiz scores*, we draw a **scatterplot**. Each student is one point on the graph, using our original 6-student sample.



The pattern moves up and to the right: students who studied more tended to score higher on the quiz. This visual impression will later match the idea of *positive correlation*.

4 Describing One Variable: Mean and Standard Deviation

Now we attach some numbers to the picture. For a single numerical variable, two basic summaries are:

- the **mean** (average) for center,
- the **standard deviation** for spread.

4.1 Sample mean

Given sample values $x_1, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

It is a statistic that estimates the population mean μ.

Running example: mean quiz score

The quiz scores are

$$y_1, \ldots, y_6 = 4, 5, 6, 7, 8, 9.$$

The sample mean is

$$\bar{y} = \frac{4+5+6+7+8+9}{6} = \frac{39}{6} = 6.5.$$

We interpret: in this sample, the average quiz score is 6.5 out of 10.

4.2 Sample standard deviation

The **sample standard deviation** s describes how spread out the data are. It comes from the sample variance s^2:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad s = \sqrt{s^2}.$$

We use $n-1$ in the denominator instead of n so that s^2 is a good estimator of the population variance σ^2 in many common situations.

Running example: standard deviation of quiz scores

For the quiz scores $4, 5, 6, 7, 8, 9$, we already found $\bar{y} = 6.5$.

Compute the deviations:

$$y_i - \bar{y} = -2.5, -1.5, -0.5, 0.5, 1.5, 2.5.$$

Square them and add:

$$(-2.5)^2+(-1.5)^2+(-0.5)^2+(0.5)^2+(1.5)^2+(2.5)^2 = 6.25+2.25+0.25+0.25+2.25+6.25 = 17.5.$$

Then

$$s_y^2 = \frac{17.5}{n-1} = \frac{17.5}{5} = 3.5, \qquad s_y = \sqrt{3.5} \approx 1.87.$$

Interpretation: quiz scores in our sample are typically about 1.9 points away from the mean of 6.5.

5 Describing Two Variables: Covariance and Correlation (Preview)

We will study linear relationships more deeply in a later week. For now we just introduce the basic ideas and symbols.

Given paired data (x_i, y_i), we can measure how X and Y move together.

5.1 Sample covariance (idea only)

The **sample covariance** is

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

- If $x_i-\bar{x}$ and $y_i-\bar{y}$ tend to have the same sign, s_{xy} is positive (positive association).
- If they tend to have opposite signs, s_{xy} is negative (negative association).

5.2 Sample correlation

The **sample correlation coefficient** r is a standardized version of the covariance:

$$r = \frac{s_{xy}}{s_x s_y},$$

where s_x and s_y are the sample standard deviations of X and Y. We always have $-1 \le r \le 1$.

- $r > 0$: positive linear association.
- $r < 0$: negative linear association.
- $|r|$ close to 1: strong linear association.
- $|r|$ near 0: weak or no linear association.

Later we will connect r directly to the slope of the least squares regression line.

6 Second Example and Review

To practice, we finish with a second example that uses the same ideas on a different context.

Example: minutes of social media per day

Suppose we ask $n = 10$ randomly selected students how many minutes they spend on social media in a typical weekday. Here are the results:

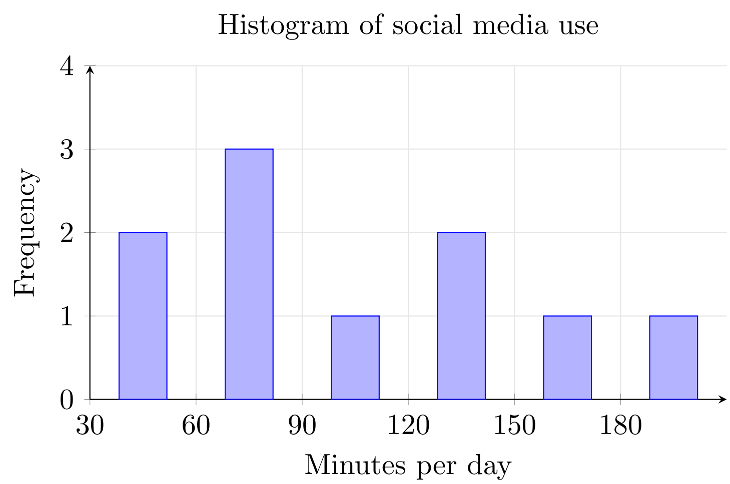
Student	1	2	3	4	5	6	7	8	9	10
Minutes x_i	30	45	60	60	75	90	120	120	150	180

6.1 Frequency table and histogram

First, make a simple frequency table by grouping into 30-minute bins:

Minutes per day	Bin	Frequency
30–59	(30–60)	2
60–89	(60–90)	3
90–119	(90–120)	1
120–149	(120–150)	2
150–179	(150–180)	1
180–209	(180–210)	1

A histogram using these bins might look like this:



We can already see that:

- many students are between about 60 and 150 minutes per day,
- a few students are much lower (around 30 minutes),
- and a few are higher (around 180 minutes).

6.2 Mean and standard deviation (with full calculation)

Step 1: Compute the mean.

$$\bar{x} = \frac{30 + 45 + 60 + 60 + 75 + 90 + 120 + 120 + 150 + 180}{10} = \frac{930}{10} = 93.$$

So in this sample, students spend an average of 93 minutes per day on social media.

Step 2: Compute deviations $x_i - \bar{x}$.

Student	x_i	$x_i - \bar{x}$
1	30	$30 - 93 = -63$
2	45	$45 - 93 = -48$
3	60	$60 - 93 = -33$
4	60	$60 - 93 = -33$
5	75	$75 - 93 = -18$
6	90	$90 - 93 = -3$
7	120	$120 - 93 = 27$
8	120	$120 - 93 = 27$
9	150	$150 - 93 = 57$
10	180	$180 - 93 = 87$

Step 3: Square the deviations and add them up.

$$\begin{aligned}
(x_1 - \bar{x})^2 &= (-63)^2 = 3969, \\
(x_2 - \bar{x})^2 &= (-48)^2 = 2304, \\
(x_3 - \bar{x})^2 &= (-33)^2 = 1089, \\
(x_4 - \bar{x})^2 &= (-33)^2 = 1089, \\
(x_5 - \bar{x})^2 &= (-18)^2 = 324, \\
(x_6 - \bar{x})^2 &= (-3)^2 = 9, \\
(x_7 - \bar{x})^2 &= (27)^2 = 729, \\
(x_8 - \bar{x})^2 &= (27)^2 = 729, \\
(x_9 - \bar{x})^2 &= (57)^2 = 3249, \\
(x_{10} - \bar{x})^2 &= (87)^2 = 7569.
\end{aligned}$$

Now add these squared deviations:

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 3969 + 2304 + 1089 + 1089 + 324 + 9 + 729 + 729 + 3249 + 7569 = 21060.$$

Step 4: Compute the sample variance and sample standard deviation.

The sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{21060}{10-1} = \frac{21060}{9} = 2340.$$

So the sample standard deviation is

$$s = \sqrt{2340} \approx 48.37.$$

We might round this to $s \approx 48.4$ minutes.

Interpretation:

- The mean $\bar{x} = 93$ tells us that, in this sample, students spend on average about 93 minutes per day on social media.
- The standard deviation $s \approx 48.4$ tells us that a typical student's time is about 48 minutes away from the mean. There is quite a lot of variation: some students are much lower than 93 minutes, while others are much higher.

Review checklist for Week 1

By the end of Week 1, you should be comfortable with:

- identifying populations and samples in a story,
- recognizing parameters (μ, σ, p, ρ) vs statistics $(\bar{x}, s, \hat{p}, r)$,
- reading and making frequency tables, bar charts, and histograms,
- computing and interpreting the sample mean $\bar{x}$ and sample standard deviation s in context,
- reading a scatterplot and recognizing a positive, negative, or weak relationship.