UNIT 2 · BLOCK 1

# Frequency Distributions

*Turning a messy pile of numbers into a picture you can read.*

## Why this matters

You have 200 students and one long column of numbers. Staring at it tells you nothing. But **count** how many fall into each group, draw a quick **graph**, and the whole story jumps out — where most people land, who's unusual, what's normal. This unit is just **counting and drawing**. No scary math at all — and it's the foundation for every average, every comparison, and every test that follows.

## By the end you can…

- ✓ Build a **frequency table** (counts) and add a **relative frequency** (%) column.
- ✓ Group numbers into **bins** when there are too many values to list.
- ✓ Choose between a **bar chart** and a **histogram** — and know why bars touch in one.
- ✓ Read a distribution's **shape**: symmetric, skewed, uniform, or bimodal.

## 1. Frequency tables — just counting

A **frequency** is simply a **count** — how many times something appears. A **frequency distribution** is the whole picture: a list of every category (or value) next to how often it occurs. List each category, write how many landed there, and you have a frequency table. Here's `class_year` for all 200 students in our dataset:

| Class year | Frequency (count) |
|---|---|
| Freshman | 56 |
| Sophomore | 51 |
| Junior | 54 |
| Senior | 39 |
| **Total** | **200** |

The counts add up to 200 because every student is counted exactly once — that total is your **n**. A close cousin is **cumulative frequency**: a running total as you move down the table (56, then 107, then 161, then 200). It answers "how many at this level *or below*?" — handy for questions like "how many students are Junior or earlier?"

## 2. Relative frequency — turning counts into %

A raw count of 56 doesn't mean much on its own — is that a lot? **Relative frequency** fixes that: divide each count by the total to get a **proportion**, then ×100 for a **percent**. This is what lets you fairly compare groups of different sizes: "28% of *our* class are Freshmen" can be compared to another class of 500 even though the raw counts differ wildly.

### The one formula

**relative frequency = count ÷ total.** So Freshmen: 56 ÷ 200 = 0.28 = **28%**. Now you can say "28% of our students are Freshmen" — much clearer than "56."

## Worked example — build the relative-frequency column

Turn the four class-year counts into percentages. Predict each step before revealing it.

**STEP 1 — GET THE TOTAL (N)**

Add the counts: 56 + 51 + 54 + 39 = `200`. Every relative frequency divides by this.

**STEP 2 — DIVIDE EACH COUNT BY THE TOTAL**

Freshman: 56 ÷ 200 = `0.28`. Sophomore: 51 ÷ 200 = 0.255. Junior: 54 ÷ 200 = 0.27. Senior: 39 ÷ 200 = 0.195.

**STEP 3 — MULTIPLY BY 100 FOR PERCENT**

0.28 → `28%`, 25.5%, 27%, 19.5%.

**STEP 4 — CHECK THE TOTAL**

28 + 25.5 + 27 + 19.5 = `100%`. ✔ If your percents don't sum to 100, you dropped or double-counted someone.

CHECK **A frequency table's raw counts add up to *n* = 200. What do its relative frequencies add up to?**

**100%. Counts sum to the sample size (200); relative frequencies are shares of the whole, so they always sum to 100% (or 1.00 as a proportion). That sum is your built-in error check.**

### 3. Grouped tables — for numbers with too many values

Categories are easy. But `commute_minutes` has values like 4, 12, 27, 33, 51… List every one and you'd have a table a mile long that hides the pattern. The fix: **group the numbers into bins** (also called **class intervals**) — equal-width ranges — and count how many fall in each. A few rules keep bins honest:

- **Equal width** — every bin spans the same range (all 15 minutes wide), or the picture lies.
- **No gaps, no overlap** — every value lands in exactly one bin. (Decide up front that, say, 15 goes in the 15–30 bin, not 0–15.)
- **About 5–8 bins** — enough to show the shape without turning into noise.

Here's the real `commute_minutes` distribution — and you can change the bin width yourself:

Notice the trade-off the slider reveals: **wide bins** give you the big picture (most commute short, a few don't); **narrow bins** reveal detail but get noisy. There's no single "right" width — you're choosing how much to zoom, and an honest analyst tries a few.

### 4. Bar chart vs. histogram — the distinction people mix up

Both are made of bars. The difference is **what's on the bottom axis** — and it follows directly from Unit 1's scales of measurement.

| | Bar chart | Histogram |
|---|---|---|
| **Use for** | Categorical data (groups) | Quantitative data (numbers) |
| **Example** | class_year, gender, diet | commute, stress, GPA |
| **Bars touch?** | No — gaps between them | Yes — bars touch |
| **Bars reorderable?** | Yes (order is arbitrary) | No (number line has a fixed order) |

**Quick test**

Counting *groups* (Freshman, Woman, Vegan)? → **bar chart**. Counting *numbers* split into ranges (0–15 min, 15–30 min)? → **histogram**. The bars touch in a histogram because the numbers flow continuously — there's no gap between "29 minutes" and "30 minutes."

CHECK **You want to graph how many students follow each `diet` (Omnivore, Vegetarian, Vegan). Which chart?**

**Bar chart. Diet is categorical (nominal) — separate groups — so the bars get gaps. A histogram is only for quantitative numbers split into ranges, like commute minutes.**

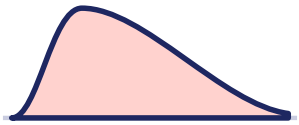
### 5. Reading the SHAPE

Once you draw a histogram, its shape tells the story at a glance — and it previews ideas we'll formalize later (it secretly tells you how the

mean and median will relate). Here's what the common shapes look like:

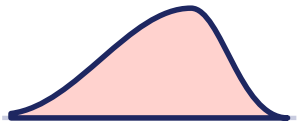
**Symmetric (unimodal)**

One central peak; the halves mirror. The classic "bell." Mean ≈ median.



**Right-skewed (+)**

Long tail to the right; most scores low. Mean > median.

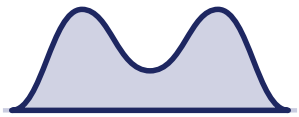


**Left-skewed (−)**

Long tail to the left; most scores high. Mean < median.

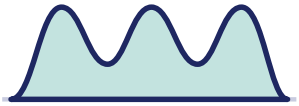
**Uniform**

Flat — every value about equally likely; no real peak.



**Bimodal**

Two peaks — often two groups mixed together.



**Multimodal**

Three or more peaks. (Two = bimodal; many = multimodal.)

- **Symmetric / bell-shaped** — two halves mirror each other; scores pile in the middle (our `stress` data). Mean ≈ median.
- **Right-skewed (positive)** — a long tail to the right; most scores low, a few high (our `commute` data). The tail drags the mean above the median.
- **Left-skewed (negative)** — the long tail points left; most scores high, a few low. The mean is pulled below the median.
- **Uniform** — every bar about the same height; no real peak.
- **Bimodal** — two separate peaks, often a sign of two groups mixed together (e.g., two class sections).

Watch, too, for **outliers** — lone values far from the pack. They're often what create a skew, and they'll matter a lot when we pick a "typical" value next unit.

**Remember the tail**

**Skew points where the tail goes.** Tail dragging right = right-skewed. The big bump is on the *opposite* side from the tail — a common exam trap.

CHECK **Most students score high on a wellbeing scale, but a few score very low, creating a long tail toward the low end (the left). The shape is…**

**Left-skewed. The tail points left, so it's left- (negative-) skewed — even though most of the data sits on the right. Name the skew by the *tail*, not the bump.**

### The biggest distribution of all

Frequency distributions aren't just about our class. Each year, roughly **83 billion land animals** are slaughtered for food — about **2,600 every second** — and the species breakdown is itself a dramatic frequency distribution: **about 90% are chickens**. Drawn as a bar chart, one bar would tower over all the rest.

Our class dataset carries a quieter version of the same story. Here's the `diet` frequency distribution of all 200 students:

| Diet | Count | Relative frequency |
|---|---|---|
| Omnivore | 136 | 68% |
| Vegetarian | 46 | 23% |
| Vegan | 18 | 9% |
| **Total** | **200** | **100%** |

Zoom out and the numbers reframe everything: meat and dairy use **83% of the world's farmland** but provide only **18% of our calories** (Poore & Nemecek, 2018). A "distribution" can carry real moral weight — and that's exactly the kind of thing we'll graph, unpack, and argue about in class.

### In Google Sheets

You don't count by hand — Sheets does it. Open the **SJNY Student Life** dataset and try:

```
=COUNTIF(C2:C201, "Freshman")        → counts how many Freshmen
56 / 200                              → relative frequency (0.28 = 28%)
```

Swap "Freshman" for each category to fill the whole table. For **bins** (number ranges), use `=COUNTIFS` with two conditions — at or above the low end, below the high end:

```
=COUNTIFS(W2:W201, ">=0",  W2:W201, "<15")   → the 0–15 bin
=COUNTIFS(W2:W201, ">=15", W2:W201, "<30")   → the 15–30 bin
```

To graph: highlight your table, then **Insert ▸ Chart**, and pick a column chart (categories) or histogram (numbers) in the Chart editor. Tip: put relative frequencies on the axis instead of counts when you want to compare two groups of different sizes.

### Stat in the Wild — chart crimes

Graphs can **lie** without a single false number — just by drawing the picture sneakily. The classic trick is a **truncated y-axis**: instead of starting at zero, the axis starts at, say, 95. A tiny gap between 96% and 98% suddenly looks *enormous*. You see it in ads ("4× more effective!"), cable-news charts, and product comparisons constantly.

**The lesson:** before you believe a bar chart, **check the y-axis**. If it doesn't start at zero, the differences are being exaggerated on purpose. A graph is an argument — make it prove its case.

### Watch out for

Counts (56) vs. percents (28%) are different things — label which you mean. Bins must be equal-width and not overlap. "Skewed right" means the *tail* is on the right even though most data sit on the left. And always check a chart's axis before trusting how big a difference looks.

## Key terms

**Frequency** · **Frequency distribution** · **Relative frequency** · **Proportion** · **Cumulative frequency** · **Class interval / bin** · **Bar chart** · **Histogram** · **Symmetric** · **Skew (left/right)** · **Uniform** · **Bimodal** · **Outlier**

## Check yourself

Ten quick true/false items with instant explanations. (Different from your Canvas quiz — real practice.)

**Sigma says:** When you're stuck on bar-vs-histogram, ask one thing — *am I counting groups or numbers?* Groups get gaps; numbers touch. That single question answers most exam items.