Graphs That Answer Statistical Questions

Lectures 3–4

QBA201 – Quantitative Business Analysis

Shi Qiu – Fall 2026

The question concerns the game's users

A gaming company wants to describe the ages of subscribers to a new PC game. The analysis should identify the age groups represented and the spread of ages.

Discuss: Would the average age alone answer both questions?

A long list hides the distribution

In the sample, ages range from 8 to 53 years.

We want to know where ages cluster, how spread out they are and whether the shape has a long tail.

A frequency distribution

Count observations within a set of nonoverlapping age intervals.

Every observation belongs to exactly one class

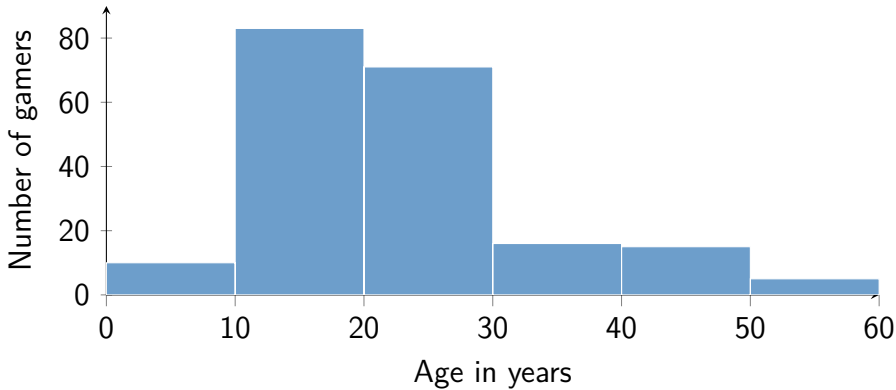
Use six ten-year classes:

$$[0, 10],\ (10, 20],\ (20, 30],\ (30, 40],\ (40, 50],\ (50, 60].$$

The right endpoint belongs to its class. For integer ages, $(10, 20]$ includes 11 through 20.

Discuss: Which class contains an age of exactly 20?

A histogram shows the grouped distribution



Turn counts into defensible statements

There are 83 gamers aged 11–20 and 71 aged 21–30.

$$\frac{83}{200} = 41.5\%, \qquad \frac{71}{200} = 35.5\%.$$

Together, these two groups account for:

$$\frac{83 + 71}{200} = 77\%.$$

Range and class width are different quantities

$$\text{Range} = 53 - 8 = 45 \text{ years}.$$

Dividing the range by a proposed class count suggests a width.

Convenient boundaries can extend beyond the minimum and maximum.

This example uses six ten-year classes, covering 0 to 60.

A histogram has numerical intervals on its axis

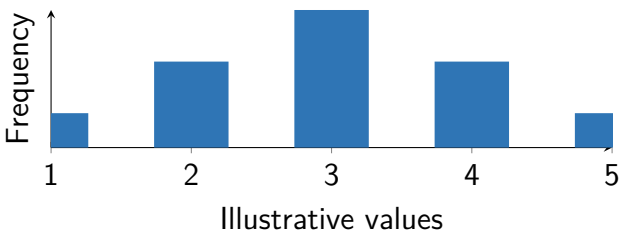
- Bars represent adjacent numerical intervals.
- Equal-width classes let bar heights represent counts.
- A bar chart of categories represents a different kind of comparison.

With unequal class widths, use frequency density so bar area represents frequency.

Discuss: Why would unequal widths distort a count-height histogram?

Symmetry concerns the two sides of the distribution

A symmetric distribution has a similar shape on each side of its center.



Skewness follows the direction of the long tail

Right-skewed

A long tail extends toward larger values. A few very large observations pull the tail to the right.

Left-skewed

A long tail extends toward smaller values.

Discuss: If most delivery times are short but a few are extremely long, which skew is plausible?

Modality describes the number of peaks

Unimodal

One main peak.

Bimodal

Two main peaks.

Two peaks can suggest a mixture of groups, but a graph alone does not identify those groups.

Discuss: Why might a histogram of all employees' commute times have two peaks?

A bell shape is a particular pattern

A bell-shaped distribution is approximately symmetric with one main central peak and thinner tails.

A histogram can help assess whether that pattern is plausible.

A roughly bell-shaped histogram is evidence about shape, not proof of an exact normal model.

Class activity: daily social-media use

Privately write your typical daily social-media time in hours.

1. Agree on a reference period, such as the past seven days.
2. Group the class responses into consistent intervals.
3. Sketch a histogram.

Discuss: Where is the main concentration? Is the distribution symmetric or skewed?

Relative frequency changes the vertical scale

$$\text{Relative frequency} = \frac{\text{class count}}{200}.$$

The first class has $10/200 = 0.05$.

The 31–40 class has $16/200 = 0.08$.

The 51–60 class has $5/200 = 0.025$.

Divide every count by the same sample total. The shape stays the same.

Cross-sections and time series

Cross-section

Different units observed at one point or period.
Example: prices at 30 shops today.

Time series

One quantity observed across successive periods.
Example: monthly petrol prices at one location.

A time plot puts time on the horizontal axis.

Removing inflation changes the question

Let P_t be a nominal price and CPI_t a price index with base 100.

$$\text{Price in base-period money} = \frac{P_t}{CPI_t} \times 100.$$

For the supplied January 1976 observation:

$$\frac{60.5}{55.8} \times 100 = 108.42.$$

A higher nominal price need not mean a higher inflation-adjusted price.

A scatterplot keeps paired measurements together

Each point records two measurements from the same unit.

$$(x_i, y_i) = (\text{size of home } i,\ \text{price of home } i).$$

Place the explanatory variable on the horizontal axis and the outcome on the vertical axis.

Axis placement does not establish causation.

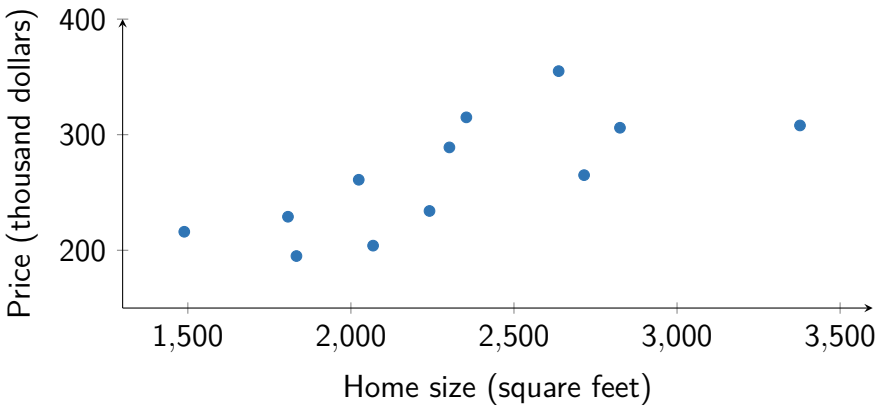
A home-size and selling-price example

The supplied sample contains 12 recently sold homes.

- x: home size, using the raw size values as square feet.
- y: selling price, in thousands of dollars.

Discuss: What pattern would you expect before looking at the points?

The home data show a positive association



Direction, form and strength

Direction

Do larger x values tend to accompany larger or smaller y values?

Form

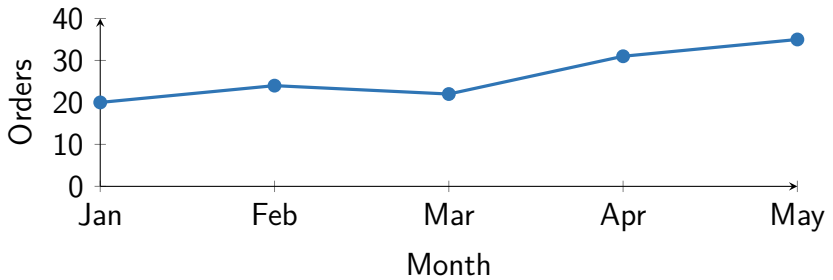
Does the pattern look approximately linear, curved or more complex?

Strength

How tightly do points follow that pattern?

Discuss: Can a strong curved relationship have a weak linear correlation?

A time plot keeps the sequence visible



Constructed example: orders generally rise, with a small fall in March.

Keep time on the horizontal axis and periods in their actual order.

Separate an observed pattern from a causal claim

The home data show a positive association between size and price.

Larger homes may also differ in location, quality, and land size.

The scatterplot alone does not isolate the causal price effect of adding one square foot.

A final check before presenting a graph

1. Do the axes identify variables and units?
2. Does the scale support a fair comparison?
3. Are the counts or percentages correct?
4. Does the written conclusion describe only what the evidence supports?

Discuss: Choose one graph from this lecture and explain its main message in one sentence.