Data Types and Categorical Comparisons

Lecture 2

QBA201 – Quantitative Business Analysis

Shi Qiu – Fall 2026

Variable, possible values and observed data

Variable

The characteristic recorded for each unit, such as a student's mark.

Possible values

Marks can range from 0 to 100 under the stated marking scale.

Observed data

The values actually recorded: $67, 74, 71, 83, 93, 55, 48$.

Discuss: Is a student ID a measurement of how much of something a student has?

Nominal data: numbers can be labels

Marital-status codes can label categories:

$$1 = \text{single}, \quad 2 = \text{married}, \quad 3 = \text{divorced}, \quad 4 = \text{widowed}.$$

The codes identify categories. They do not measure an amount.

Discuss: Would using codes 10, 20, 30 and 40 change the people in each group?

Count categories and compare proportions. Do not average arbitrary codes.

Ordinal data: order without a known distance

$$\text{poor} < \text{fair} < \text{good} < \text{very good} < \text{excellent}.$$

The categories have an order.

The distance from poor to fair need not equal the distance from good to very good.

Discuss: Does rating 4 represent twice the satisfaction of rating 2?

Quantitative data have interpretable units

- AED80 is AED20 greater than AED60.
- A 40-minute delivery is 10 minutes longer than a 30-minute delivery.
- A purchase count takes whole-number values.

State the unit as well as the numerical value.

Interval and ratio scales

Interval scale

Differences are meaningful. Zero is an arbitrary reference.
20°C is not twice as hot as 10°C.

Ratio scale

Differences and ratios are meaningful because zero represents absence.
20 kg is twice 10 kg.

Grouping data can discard information

Exact age records more detail than age bands.

$$\text{Exact age } 24 \quad \longmapsto \quad \text{age band } 20\text{–}29.$$

Age bands record ordering but cannot recover each exact age.

More detailed measurements can be grouped. Grouped data cannot generally recover the original detail.

Check: when is a numerical average meaningful?

A spreadsheet contains 1, 2 and 3.

Discuss: What must you know before interpreting their mean?

Possible meanings

Purchase counts, satisfaction codes or branch labels.

The variable definition determines which calculations mean something.

From counts to relative frequencies

Let n_j be the count in category j, and n the total.

$$\text{Relative frequency}_j = \frac{n_j}{n}, \qquad \text{Percentage}_j = 100\frac{n_j}{n}.$$

For mutually exclusive, exhaustive categories:

$$\sum_j n_j = n, \qquad \sum_j \frac{n_j}{n} = 1.$$

A percentage needs a denominator.

The frequency distribution

Work status	Count	Percent
Employed	978	45.38
Self-employed	108	5.01
Unemployed	155	7.19
Full-time student	84	3.90
Retired	472	21.90
School	97	4.50
Home and family	97	4.50
Long-term sick or disabled	112	5.20
Other	52	2.41
Total	2,155	100.00

One row, calculated carefully

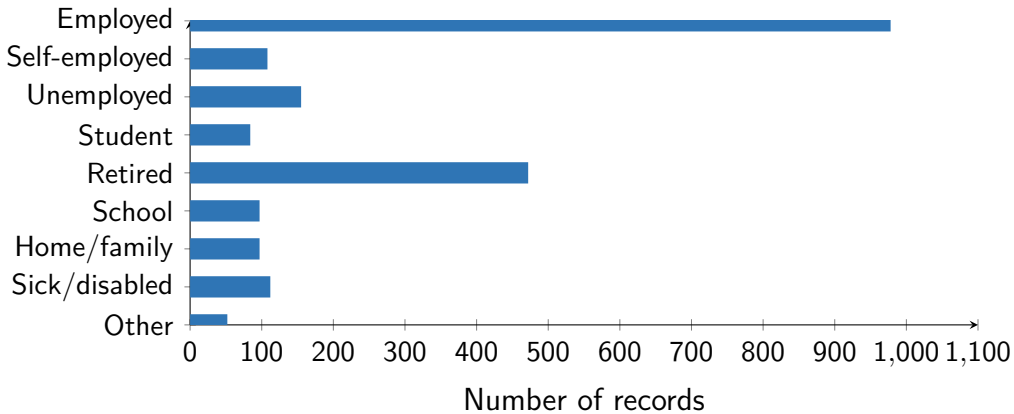
For employed respondents, $n_j = 978$ and $n = 2,155$.

$$\frac{978}{2,155} = 0.4538, \qquad 100\left(\frac{978}{2,155}\right) = 45.38\%.$$

Interpretation

About 45.38% of the records in this dataset are classified as employed.

A bar chart compares category counts



A pie chart represents shares of one total

The full circle represents 100%.

$$\text{Sector angle} = 360^\circ \left(\frac{n_j}{n} \right).$$

For employed respondents:

$$360^\circ \left(\frac{978}{2,155} \right) = 163.38^\circ.$$

Discuss: Why must categories be nonoverlapping parts of the same whole?

Bar chart or pie chart?

Bar chart

Aligned lengths make small differences easier to compare.
Counts or percentages can determine the bar heights.

Pie chart

Shows composition of one whole.
Many small slices make precise comparison difficult.

Choose the display that makes the intended comparison easiest to read.

Two categorical variables need a two-way table

A cross-classification table counts each combination of categories.

- Rows identify occupation.
- Columns identify newspaper.
- Each cell counts readers in that combination.

Keep the two labels for each reader together.

The readership question

The textbook survey records one newspaper choice and one occupation category per reader.

Newspapers: G&M, Post, Star and Sun.

Occupations: blue-collar, white-collar and professional.

Discuss: Does the mix of newspapers look the same across occupations?

The cross-classification counts

Occupation	G&M	Post	Star	Sun	Total
Blue-collar	27	18	38	37	120
White-collar	29	43	21	15	108
Professional	33	51	22	20	126
Total	89	112	81	72	354

Discuss: What denominator gives the share of blue-collar readers choosing Post?

Compare percentages within each occupation

Occupation	G&M	Post	Star	Sun
Blue-collar	22.50%	15.00%	31.67%	30.83%
White-collar	26.85%	39.81%	19.44%	13.89%
Professional	26.19%	40.48%	17.46%	15.87%

For blue-collar readers choosing Post: $18/120 = 0.15 = 15\%$.

Divide by the row total to compare newspaper choices within occupation.

Compare composition rather than group size

Raw counts can differ because the occupation groups contain different numbers of respondents.

Within-group percentages provide a common scale.

$$\text{Post shares: } 15.00\%, \quad 39.81\%, \quad 40.48\%.$$

Discuss: Which two occupation groups have the most similar Post readership shares?

An observed association is a description

White-collar and professional readers have similar readership patterns. The blue-collar group has a different mix in this sample.

- This describes an association.
- It does not establish that occupation causes newspaper choice.
- Population conclusions need a defensible sampling and inference method.