ECOM5005 Tutorial 1

Topic 1: Defining and Collecting Data

Curtin University

School of Accounting, Economics and Finance

Aligned with Topic 1: Defining and Collecting Data

Preparation and learning objectives

Preparation: Attempt every question before class. Where a question is open-ended, prepare a defensible position and identify the assumptions on which your answer depends.

- Classify variables by data type and level of measurement.
- Evaluate the quality of data sources and identify potential biases.
- Distinguish populations, samples, parameters and statistics.
- Apply normalisation, inflation adjustment and index-number reasoning.
- Select an appropriate probability-sampling method for a real business problem.

Part 1 - Pearson MyLab

Please use the instructions in the unit Welcome module to access Pearson MyLab. Complete the candidate agreement before the tutorial where possible.

Part 2 - Short-answer and applied questions

Question 1. Data type and level of measurement: beverage choice

A fast-food restaurant records each drink sold as one of three types: soft drink, fruit juice or coffee.

(a) Explain why drink type is a categorical variable.

(b) Explain why drink type is measured on a nominal scale.

Question 2. Data type and level of measurement: download time

Suppose that you measure the time required to download a digital audio file from the internet.

(a) Explain why download time is a numerical variable and state whether it is discrete or continuous.

(b) Explain why download time is measured on a ratio scale.

Question 3. Classifying home-loan application data

Metro Home Loans collects the following information from Andrew and Fiona Chen. Classify each response by data type, and where appropriate, as discrete or continuous. Also state the level of measurement.

Item	Recorded response	Your classification
Monthly household expenses	$2,056	
Number of dependants	2	
Annual family income	$105,000	
Marital status	Married	

Question 4. Practicality of a simple random sample

A national organisation plans to conduct personal, face-to-face interviews rather than online or telephone interviews. Explain why a simple random sample of individuals may be less practical than another probability-sampling method. Name a more practical alternative and discuss one trade-off.

Question 5. Evaluating a survey claim

A survey reports that “the vast majority of university students own their own personal computer.” What information would you require before accepting this result as credible? Organise your response under the headings population and sampling, measurement, fieldwork, and statistical reporting.

Question 6. Online restaurant ratings and self-selection

An online restaurant-booking platform rewards diners with points for submitting ratings. A restaurant at The Rocks in Sydney receives the following scores: Recommended 8.7; Food 8.5; Service 8.7; Value for money 7.8; Atmosphere 8.4. Explain why these scores may differ from ratings obtained from a probability sample of all diners. Identify at least four potential sources of distortion.

Question 7. Interpreting a “majority” statement

An ECOM5005 tutorial has 40 students enrolled, and 17 are recorded as male. A student states, “A majority of students are female.” Assume that 17 and 40 are recorded correctly. Is the statement necessarily valid? Discuss the meanings and assumptions embedded in the words majority, students and female.

Question 8. Designing a feasible business-analytics project

Propose a non-trivial business or policy question for which evidence must realistically be obtained from a sample rather than a census. Your proposal must specify:

- the decision to be supported;
- the target population and sampling frame;
- the key outcome and explanatory variables;
- a probability-sampling method;
- two likely sources of bias or error; and
- one ethical or privacy issue.

Question 9. Integrated case: FlexiPay customer analytics

FlexiPay is a national buy-now-pay-later provider. It is considering whether to increase credit limits for customers aged 18-30. Its database contains the variables shown below. To measure satisfaction, FlexiPay sends an in-app survey only to customers who made a purchase in the previous 30 days. Thirty-eight per cent respond. The survey asks respondents to rate the statement, "FlexiPay's convenient and responsible service deserves a high rating," from 1 (strongly disagree) to 5 (strongly agree).

Variable	Example / format
Annual income	Australian dollars
Late payments in previous 12 months	Count
Employment type	Full-time, part-time, casual, self-employed, unemployed
Satisfaction score	1 to 5
Customer complaint	Free-text comment
App-login time	Timestamp

(a) Classify each variable by data type, and where applicable, level of measurement. State any reasonable convention you adopt.

(b) Distinguish the structured and unstructured data in the case. Which data are primary and which are secondary for the satisfaction study?

(c) Identify and explain possible coverage/selection bias, non-response bias and response bias.

(d) Design a stratified probability sample that would better represent the relevant customer population. Define the strata and explain how respondents would be selected.

(e) Give one descriptive, one diagnostic, one predictive and one prescriptive analytics question that FlexiPay could investigate.

Question 10. Integrated case: inflation, normalisation and index numbers

FreshCart operates supermarket-delivery branches in Perth and Sydney. Management is comparing performance between 2021 and 2025. Each city's CPI is set equal to 100 in 2021.

Branch	2021 revenue ($m)	2021 transactions (000s)	2021 CPI	2025 revenue ($m)	2025 transactions (000s)	2025 CPI
Perth	12.00	240	100	17.64	280	112
Sydney	18.00	300	100	25.65	360	114

(a) Calculate nominal revenue growth from 2021 to 2025 for each branch.

(b) Convert each branch's 2025 nominal revenue into 2021 dollars using Real value = Nominal value / (CPI/100).

(c) Calculate real revenue per transaction in 2021 and 2025, and the percentage change for each branch.

(d) Which branch shows the stronger improvement? Explain why the answer depends on whether the objective is absolute revenue, real growth or real revenue per transaction.

(e) Can management conclude that Sydney had a higher cost of living in 2025 simply because its CPI was 114 and Perth's was 112? Explain carefully.

Part 3 - Data exercise: analysing WAM by recorded gender

Purpose. Use the supplied Excel dataset (First, open the link titled “survey questions” and the link titled “survey results.xlsx” for Topic 1. The latter contains answers provided by 62 people and the former contains the questions.) to complete a reproducible descriptive analysis of WAM for respondents recorded as Female and Male. The analysis is descriptive: it summarises this sample and does not establish causation.

Important: Use the desktop version of Microsoft Excel. Keep the original DATA worksheet unchanged, work on a copy, and save the file as an Excel Workbook (*.xlsx). Do not delete an observation merely because it appears unusual.

Files and required workbook structure

Open the supplied file survey results.xlsx. The source data occupy DATA!A1:N63: row 1 contains headers and rows 2-63 contain respondent records.

Worksheet	Required content	Purpose
DATA	Retain the original data without deleting or reordering records.	Provides an unchanged source sheet.
Sorted Data	Copy the complete data block and sort the full rows by Gender.	Creates continuous Female and Male WAM ranges.
Part 3 Analysis	Dataset checks, descriptive-statistics comparison, common-bin frequency table, key calculations and final charts.	Presents the final analysis in one place.

Task 1. Open the dataset and save a working copy

(a) Open survey results.xlsx in the desktop version of Excel.

(b) Select File > Save As and save the working file as Tutorial 1 Part 3 - Your Name.xlsx.

(c) Confirm that the file type is Excel Workbook (*.xlsx). Do not overwrite the original file.

(d) Keep the worksheet DATA unchanged. Create two additional worksheets named Sorted Data and Part 3 Analysis.

Task 2. Audit the dataset before analysing it

On Part 3 Analysis, create a small dataset-check table. Enter the formulas below and record the results. Use the formulas as written so that another analyst can reproduce your checks.

Check	Excel formula	What the result establishes
Non-blank Gender entries	=COUNTA(DATA!A2:A63)	Number of records with a recorded gender.
Numeric WAM observations	=COUNT(DATA!F2:F63)	Number of usable numeric WAM values.
Missing WAM values	=COUNTBLANK(DATA!F2:F63)	Whether any WAM values are blank.
Female observations	=COUNTIF(DATA!A2:A63,"Female")	Size of the Female group.
Male observations	=COUNTIF(DATA!A2:A63,"Male")	Size of the Male group.

(a) Inspect the Gender column for unexpected spellings, leading or trailing spaces, missing categories or categories other than Female and Male.

(b) Inspect WAM for blanks, text stored as numbers and values outside the possible range 0 to 100.

(c) If a WAM value appears unusually low or high, state what you would verify before changing or deleting it.

Task 3. Create the Sorted Data worksheet

(a) Copy DATA!A1:N63 to the worksheet Sorted Data. Do not sort the original DATA sheet.

(b) Select the entire copied data block, choose Data > Sort, tick My data has headers, sort by Gender, and use A to Z order.

(c) Confirm the row ranges occupied by Female and Male records after sorting, and record those row ranges in a small note block on Sorted Data.

(d) Record the column containing WAM.

(e) Explain why sorting only the Gender column, rather than the complete rows, would corrupt the dataset.

Task 4. Enable the Analysis ToolPak

(a) Windows: File > Options > Add-ins > Manage: Excel Add-ins > Go > tick Analysis ToolPak > OK.

(b) Mac: Tools > Excel Add-ins > tick Analysis ToolPak > OK.

(c) Confirm that Data Analysis appears on the Data tab. Record any installation difficulty for discussion in the tutorial.

Task 5. Produce descriptive statistics for Female and Male WAM

Use Data > Data Analysis > Descriptive Statistics. Analyse the Female and Male WAM ranges separately. For each run, choose Grouped By: Columns, leave Labels in first row unticked, tick Summary statistics, and tick Confidence Level for Mean at 95%.

(a) Use Sorted Data!F2:F34 for the Female analysis.

(b) Use Sorted Data!F35:F63 for the Male analysis.

(c) Place the two outputs side by side, or use them to build one clean comparison table on Part 3 Analysis.

(d) Your comparison table must include every measure listed below.

Required ToolPak measure	Additional check or calculation	Reason for inclusion
Mean; Standard Error; Median	Calculate Female minus Male differences for the mean and median.	Compares the centre of the samples.
Mode	Inspect the raw data for no mode or multiple modes; do not rely only on the single value printed by ToolPak.	Excel may report only one tied mode.
Standard Deviation; Sample Variance	Calculate Female standard deviation minus Male standard deviation.	Compares dispersion.
Kurtosis; Skewness	Use together with the histograms.	Describes distributional shape.
Range; Minimum; Maximum	Identify values that merit checking.	Shows overall spread and extremes.
Sum; Count	Check that group counts reconcile with Task 2.	Detects incorrect ranges.
Confidence Level (95.0%)	Calculate mean - confidence amount and mean + confidence amount.	Creates the interval endpoints reported from ToolPak output.

(e) Explain why the 95% intervals should be interpreted cautiously if the respondents were not selected through a defensible probability-sampling design.

Task 6. Construct identical WAM bins and common-bin frequency tables

Create one common set of bin upper limits so that the two distributions are compared using exactly the same intervals. Excel treats each bin value as an upper limit.

(a) On Part 3 Analysis, enter the bin upper limits 10, 20, 30, ..., 100.

(b) Create readable interval labels: WAM ≤ 10, 10 < WAM ≤ 20, 20 < WAM ≤ 30, ..., 90 < WAM ≤ 100.

(c) Use Data > Data Analysis > Histogram for Female WAM with Input Range Sorted Data!F2:F34 and the common bin range.

(d) Repeat for Male WAM with Input Range Sorted Data!F35:F63 and exactly the same bin range.

(e) Present the final output as one clean table with four columns: Bin upper limit, WAM interval, Female frequency and Male frequency.

(f) Add the frequencies in each group and confirm that each total equals the corresponding group count from Task 2. Explain why this is an important error check.

Task 7. Create the final three figures

(a) Select the readable WAM interval labels and both frequency columns. Insert a Clustered Column chart comparing Female and Male frequencies in the same figure.

(b) Use the title WAM distributions using identical 10-mark bins. Label the horizontal axis WAM interval and the vertical axis Number of students. Include a legend.

(c) Set the vertical axis to Minimum = 0, Maximum = 12 and Major unit = 2.

(d) Create a separate Female chart titled Female WAM distribution (n = 33) and a separate Male chart titled Male WAM distribution (n = 29).

(e) Use exactly the same interval labels and vertical-axis settings for both separate charts. Place all three charts neatly on Part 3 Analysis without covering the tables.

(f) Explain why different bins or different axis scales could create a misleading visual comparison.

Task 8. Interpret the final tables and figures

Answer each part in complete sentences. Support every numerical claim with the output in your workbook.

(a) Dataset composition. Report the total number of usable WAM observations, the size of each recorded-gender group and the number of missing WAM values. Is the sample reasonably balanced?

(b) Centre. Report and compare the means and medians. Calculate Female minus Male for each measure. Are the differences large relative to the within-group standard deviations?

(c) Variability. Which group is more dispersed? Support your answer with the sample standard deviation, range, minimum and maximum.

(d) Shape. Use both the histograms and skewness values. Is either distribution strongly skewed? Explain.

(e) Unusual observations. Identify any very low or very high WAM values that merit checking. Why should an unusual value not be deleted automatically?

(f) Modes. State whether each group has no mode, one mode or multiple modes. Explain the limitation of relying only on the mode displayed by the Analysis ToolPak.

(g) Frequency-table reading. Identify the most common 10-mark interval for each group. Compare the groups in $70 < \text{WAM} \leq 80$ and identify intervals containing only one observation.

(h) Confidence intervals. Report the interval endpoints calculated from the ToolPak confidence amounts. What assumptions or sampling limitations restrict generalisation beyond this sample?

(i) Overall conclusion. Write a concise descriptive conclusion comparing centre and spread. Can these data establish that recorded gender causes a difference in academic performance? Discuss possible confounding variables and the distinction between descriptive and causal claims.

Task 9. Required submission

Submit one Excel workbook. Do not submit charts as screenshots without the underlying tables and calculations. The workbook must contain:

1. the unchanged DATA worksheet;
2. the complete Sorted Data worksheet, sorted by Gender without breaking row integrity;
3. the Part 3 Analysis worksheet with the dataset-check formulas and results;
4. Female and Male descriptive-statistics outputs or a complete clean comparison table;
5. the 95% interval endpoint calculations and the Female-minus-Male differences requested in Task 5;
6. the common-bin frequency table using upper limits 10, 20, ..., 100;
7. one combined chart and two separate charts using identical bins and the required vertical scale; and
8. written responses to every interpretation question in Task 8.

Final checking rule: the Female and Male descriptive-statistics counts must agree with the dataset-check counts, and the two frequency-table totals must agree with the same counts. A mismatch indicates that an input range, row or bin range has been selected incorrectly.