ECOM5005 Tutorial 1 - Solutions

Topic 1: Defining and Collecting Data

Curtin University

School of Accounting, Economics and Finance

Aligned with Topic 1: Defining and Collecting Data

Part 2 - Solutions

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

Solution

(a) Drink type is categorical because each observation is a label that places a sale into a group: soft drink, fruit juice or coffee. Arithmetic operations such as adding or averaging these labels have no substantive meaning.

(b) The variable is nominal because the categories have no inherent ranking. Coffee is not "higher" or "lower" than fruit juice. Numeric codes could be assigned for storage, such as 1, 2 and 3, but those codes would remain labels and would not create an order.

Key principle: A numeric code does not automatically make a variable numerical. The meaning of the values, rather than their appearance, determines the data type.

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

Solution

(a) Download time is numerical because it measures an amount of time. Conceptually it is continuous: a download may take 8.2, 8.21 or 8.213 seconds, subject only to the precision of the measuring device.

(b) It is ratio-scaled because equal differences are meaningful and there is a genuine zero. Zero seconds represents no elapsed download time. Ratios are also meaningful: 20 seconds is twice as long as 10 seconds.

Measurement caveat: If software records only whole seconds, the stored values appear discrete, but the underlying variable remains continuous.

Question 3. Classifying home-loan application data

Solution

Item	Data type	Discrete / continuous	Level	Reasoning
Monthly household expenses	Numerical	Usually treated as continuous	Ratio	A true zero is meaningful and ratios are interpretable. Recorded dollars/cents are technically discrete, but financial models generally treat the underlying amount as continuous.
Number of dependants	Numerical	Discrete	Ratio	It is a count taking non-negative whole numbers. Zero means no dependants and four is twice two.
Annual family income	Numerical	Usually treated as continuous	Ratio	A true zero exists and both differences and ratios are meaningful. Recorded currency is rounded to monetary units.
Marital status	Categorical	Not applicable	Nominal	The response is a category with no natural ordering.

Question 4. Practicality of a simple random sample

Solution

A simple random sample may scatter selected individuals across the entire country. Face-to-face interviewing would then require substantial travel, interviewer time, scheduling and accommodation costs. The method may be statistically attractive but operationally inefficient.

- A more practical alternative is cluster sampling. The organisation could randomly select a number of workplaces, suburbs or branches and interview all eligible people, or a random sample of people, within the selected clusters.
- The trade-off is precision. People within a cluster are often similar to one another, so a cluster sample may contain less independent information than a simple random sample of the same size. The analysis must also account for the clustered design.
- Stratified sampling could also be useful when representation of regions or business units is important, but it does not necessarily eliminate travel because samples are taken from every stratum.

Question 5. Evaluating a survey claim

Solution

The claim cannot be assessed from the headline alone. The following information is required.

Population and sampling

- What precisely is the target population: all university students in Australia, one university, domestic students, currently enrolled students, or only students attending classes?
- What sampling frame was used, and which students were excluded from it?
- Was the sample selected using a probability method? Was it simple random, stratified, systematic or cluster sampling?
- What was the sample size and, where relevant, the design effect or margin of error?

Measurement

- How was “own” defined? Does a shared family computer count? Does a tablet count? Must the device be currently operational?
- What exact question and response options were used? Was the wording neutral and unambiguous?
- How was “vast majority” defined? It is not a precise statistical threshold.

Fieldwork

- When and how were students contacted? Online-only recruitment could under-cover students with limited internet access.
- What was the response rate, and did respondents differ from non-respondents?
- Who funded and administered the survey, and could the sponsor have an interest in the result?

Statistical reporting

- Was a confidence interval reported rather than only a sample percentage?
- Were weighting, missing responses and duplicate responses handled transparently?
- Can the result be generalised beyond the institutions and time period actually studied?

Only after these issues are examined can we judge whether the sample statistic is a credible estimate of the relevant population proportion.

Question 6. Online restaurant ratings and self-selection

Solution

- **Voluntary-response bias:** Diners choose whether to submit a rating. People with extremely positive or negative experiences may be more motivated to respond than typical diners.
- **Coverage bias:** The platform includes only diners who booked through the platform or were aware of the rating system. Walk-ins, telephone bookings and diners using other platforms may be excluded.
- **Incentive effects:** Reward points may encourage frequent platform users to respond, so the rating distribution can over-represent loyalty-program participants.
- **Repeat observations:** The same person may rate the restaurant after multiple visits. The scores may therefore represent visits rather than independent diners.
- **Response and measurement bias:** The wording, scale, timing of the request and desire to earn points may affect the ratings. Respondents may also interpret "value" or "atmosphere" differently.
- **Non-response bias:** Even among eligible platform diners, respondents may systematically differ from non-respondents.

A probability sample drawn from the complete set of diners over a defined period would reduce selection bias because every eligible diner would have a known, non-zero chance of selection. It would not automatically remove non-response or response bias; follow-up procedures and neutral measurement would still be required.

Question 7. Interpreting a "majority" statement

Solution

If the variable has exactly two exhaustive categories, male and female, and all 40 enrolled students have a valid recorded category, then the number recorded as female is:

$$\text{Female count} = 40 - 17 = 23$$

$$\text{Female proportion} = 23 / 40 = 0.575 = 57.5\%$$

Under those assumptions, more than half of the enrolled class is recorded as female, so the statement is numerically true. It is not necessarily valid without the assumptions, however.

- "Majority" normally means strictly more than 50%, not merely the largest category. If there were three categories, the largest group could be a plurality without being a majority.
- "Students" could mean all enrolled students, students attending the particular tutorial, students present on the day, or all students in ECOM5005. The denominator changes with the definition.
- The inference assumes that every non-male record is female. Contemporary gender data may include non-binary, self-described, missing or prefer-not-to-say categories.
- The statement should distinguish a recorded administrative category from a person's identity and should use respectful, current terminology.
- The timing matters: enrolment and attendance can change during semester.

A defensible statement would be: "Of the 40 students currently enrolled in this tutorial, 23, or 57.5%, are recorded as female, assuming the administrative gender variable is complete and contains only male and female categories."

Question 8. Designing a feasible business-analytics project

Illustrative model answer

Project question: Does hybrid working improve employee retention intentions and self-reported productivity in a national professional-services firm?

Design element	Model response
Decision	Whether to retain, expand or redesign the hybrid-work policy.
Target population	All current employees of the firm in Australia. The HR employee register is the sampling frame.
Outcome variables	Intention to remain for the next 12 months (yes/no/unsure), self-rated productivity on a 1-5 ordinal scale, and days absent in the previous quarter.
Explanatory variables	Number of remote-work days, occupation, business unit, tenure, employment status, office location and caring responsibilities.
Sampling method	Use stratified random sampling by business unit and office location, with possible oversampling of small groups. Select employees randomly within each stratum and apply weights if sampling fractions differ.
Potential errors	Non-response bias if dissatisfied employees are more or less likely to respond; response bias if employees fear that managers can identify them; coverage error if contractors are excluded although the policy affects them.
Ethics and privacy	Collect the minimum necessary data, separate identifiers from responses, report only aggregated results, avoid small cells that could identify individuals, and explain voluntary participation.
Limit on causal interpretation	The design is observational. Employees who choose more remote days may differ systematically from others. A correlation cannot by itself establish that hybrid work caused higher productivity or retention intentions.

Question 9. Integrated case: FlexiPay customer analytics

Solution

(a) Variable classification

Variable	Type	Discrete / continuous	Level	Comments
Annual income	Numerical	Usually continuous	Ratio	Zero has a meaningful interpretation and ratios are meaningful. Recorded currency is rounded.
Late payments	Numerical	Discrete	Ratio	A non-negative count with a meaningful zero.
Employment type	Categorical	N/A	Nominal	Categories have no universal ranking.
Satisfaction score 1-5	Categorical / ordered	N/A	Ordinal	The categories are ordered, but equal numerical gaps cannot be assumed.
Complaint text	Qualitative	N/A	Not a classical scale	Unstructured text; it may later be coded into categories.
App-login time	Numerical date/time	Continuous for elapsed time	Interval for clock/calendar time	Differences are meaningful; the calendar origin is arbitrary. Elapsed duration would be ratio-scaled.

(b) Structure and source

- The database fields, such as income, late-payment count, employment type, satisfaction score and timestamps, are structured because they follow a predefined row-and-column schema.
- The free-text complaint is unstructured. It may be transformed into structured variables through coding, sentiment analysis or topic classification, but the original text remains unstructured.
- For the satisfaction study, newly collected survey responses are primary data because FlexiPay collects them for the present research purpose.
- Existing transaction, credit and login records are secondary data for this particular study because they were originally generated for operational purposes, even though FlexiPay owns them.

(c) Potential biases

- Coverage/selection bias: only recent purchasers receive the survey. Inactive customers, customers who stopped using FlexiPay after a poor experience, and customers without current app access have zero probability of selection. The survey therefore cannot represent all customers aged 18-30.
- Non-response bias: the 62% who do not respond may have systematically different satisfaction, repayment problems or digital engagement. A low response rate is not automatically biased, but it creates a serious risk when response is related to the outcome.
- Response bias: the statement is leading because it describes the service as convenient and responsible and asks whether it deserves a high rating. Respondents may also provide socially desirable answers or fear that negative answers could affect their account.

(d) Improved stratified design

Define the target population as all active and recently inactive FlexiPay customers aged 18-30 during a specified reference date. Construct a sampling frame from the customer register. A defensible stratification could cross:

- age group: 18-21, 22-25 and 26-30;
- risk/repayment status: no late payment, one late payment, two or more late payments; and
- engagement status: recent purchaser, active without a recent purchase, and recently inactive/closed account where contact is permitted.

Randomly sample customers within each stratum. Oversample small but important high-risk or inactive strata, then use survey weights to recover population-representative estimates. Use neutral wording, multiple contact modes where lawful, reminders, clear confidentiality protections and a response-rate report by stratum.

(e) Four analytics types

- **Descriptive:** What proportion of customers in each risk tier made at least one late payment in the previous 12 months?
- **Diagnostic:** Which customer characteristics are associated with late payments or low satisfaction?
- **Predictive:** What is the predicted probability that a customer will become 30 days past due after a credit-limit increase?
- **Prescriptive:** For which customers, and by how much, should credit limits be changed to maximise expected customer value subject to a default-risk constraint?

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

Solution

(a) Nominal revenue growth

$$\text{Growth rate} = (\text{2025 revenue - 2021 revenue}) \,/\, \text{2021 revenue x } 100\%$$

- Perth: (17.64 - 12.00) / 12.00 x 100% = 47.0%.
- Sydney: (25.65 - 18.00) / 18.00 x 100% = 42.5%.

On nominal revenue growth alone, Perth appears stronger. This comparison mixes changes in prices and quantities.

(b) 2025 revenue in 2021 dollars

$$\text{Real 2025 revenue} = \text{Nominal 2025 revenue} \,/\, (\text{CPI} \,/\, 100)$$

- Perth: $17.64m / 1.12 = $15.75m in 2021 dollars.
- Sydney: $25.65m / 1.14 = $22.50m in 2021 dollars.
- Real revenue growth is therefore 31.25% in Perth and 25.0% in Sydney.

(c) Real revenue per transaction

Branch	2021 real revenue/transaction	2025 nominal revenue/transaction	2025 real revenue/transaction	Real change	Calculation
Perth	$12.00m / 240,000 = $50.00	$17.64m / 280,000 = $63.00	$15.75m / 280,000 = $56.25	12.5%	($56.25-$50)/$50
Sydney	$18.00m / 300,000 = $60.00	$25.65m / 360,000 = $71.25	$22.50m / 360,000 = $62.50	4.17%	($62.50-$60)/$60

(d) Which branch improved more?

- Absolute 2025 revenue is higher in Sydney: $25.65m nominal and $22.50m real, because the Sydney branch is larger.
- Nominal and real revenue growth are higher in Perth: 47.0% nominal and 31.25% real, compared with 42.5% and 25.0% in Sydney.
- Real revenue per transaction is still higher in Sydney in 2025: $62.50 versus $56.25. However, Perth achieved the larger improvement in this efficiency/intensity measure: 12.5% versus 4.17%.

There is no single correct ranking until management states the decision objective. Normalisation makes the metric consistent with that objective.

(e) Interpreting the CPI levels

No. The CPI values show percentage changes relative to each city's own base period. Sydney's value of 114 means its indexed basket price rose by 14% from its 2021 base, while Perth's value of 112 means a 12% increase from its base. The values do not, by themselves, reveal the absolute cost of the basket in either

city. A city can have a lower index but a higher absolute price level if it began from a higher base. A direct cost-of-living comparison requires a common basket and a price-level comparison designed for that purpose.

Part 3 - Data exercise: Excel solution

Purpose of this solution

This guide is written for students who have very limited statistics knowledge and little experience using Microsoft Excel. Follow the tasks in order. Every important click, formula, output and error check is shown. The analysis is descriptive: it summarises the supplied sample and does not prove that recorded gender causes differences in WAM.

Final numerical answer at a glance

Item	Final result
Usable WAM observations	62
Missing WAM values	0
Female / Male counts	33 / 29
Female / Male mean WAM	64.715 / 63.059
Female − Male mean difference	1.657 marks
Female / Male median WAM	64.9 / 62.4
Female − Male median difference	2.5 marks
Female / Male standard deviation	13.717 / 11.960
Main descriptive conclusion	Female centre is slightly higher; Female spread is also wider; differences are small relative to within-group variation.

How to use this solution

Complete the tasks in numerical order. Do not jump directly to the charts, because every later output depends on the dataset audit and the correct Gender sort.

- Use the desktop version of Microsoft Excel. The browser version may not include the Analysis ToolPak or may display menus differently.
- Keep the original DATA worksheet unchanged. Work only on Sorted Data and Part 3 Analysis.
- A formula begins with an equals sign (=). Type the formula exactly, then press Enter.
- Cell references such as DATA!A2:A63 mean “cells A2 through A63 on the DATA worksheet”.
- The completed workbook supplied with this solution contains all formulas, final tables and charts so that you can compare your work.

Important statistical idea

A larger mean does not automatically imply a meaningful or causal difference. Always compare the difference in means with the amount of variation within each group, and remember how the sample was obtained.

Task 1 — Open the dataset and save a working copy

The purpose of this task is to protect the original file. If you make a mistake, you must still have an unchanged source dataset.

1. **Open the file.** Locate survey results.xlsx in File Explorer (Windows) or Finder (Mac), then double-click it. Make sure it opens in the desktop Excel application.
2. **Immediately create a working copy.** Choose File > Save As. Select a folder you can find again, such as your ECOM5005 folder.
3. **Enter the required filename.** Use Tutorial 1 Part 3 - Your Name.xlsx. Replace “Your Name” with your own name.
4. **Check the file type.** The Save as type / File Format must be Excel Workbook (*.xlsx). Do not use CSV because CSV cannot preserve multiple worksheets, formulas or charts.
5. **Create the two required worksheets.** At the bottom of the workbook, click the + button twice. Double-click each new sheet tab and rename the sheets Sorted Data and Part 3 Analysis.
6. **Confirm the sheet names.** The tabs should now read DATA, Sorted Data and Part 3 Analysis. Do not rename or modify DATA.

Checkpoint before moving on

You should have three worksheets. DATA still contains the original block A1:N63. Sorted Data and Part 3 Analysis are blank.

Common mistake to avoid

Do not overwrite the original survey results.xlsx. Do not save as .csv. Do not begin deleting “unusual” observations.

Task 2 — Audit the dataset before analysing it

An audit is a set of basic checks performed before analysis. It detects missing records, incorrect ranges and impossible values. Skipping this stage is one of the most common causes of incorrect statistics.

2.1 Create the dataset-check table

1. **Go to Part 3 Analysis.** Click the Part 3 Analysis sheet tab.
2. **Create four headings.** In A3 type Check; in B3 type Excel formula; in C3 type Result; and in D3 type What the result establishes.
3. **Enter the five checks shown below.** For each row, type the formula in the Result column. After pressing Enter, Excel replaces the displayed formula with its numerical result. Click the cell again and look at the Formula Bar to confirm the formula remains stored.

Check	Formula to type	Expected result	Meaning
Non-blank Gender entries	=COUNTA(DATA!A2:A63)	62	There are 62 respondent records with a recorded Gender.
Numeric WAM observations	=COUNT(DATA!F2:F63)	62	All 62 WAM entries are recognised as numeric.
Missing WAM values	=COUNTBLANK(DATA!F2:F63)	0	There are no blank WAM cells.
Female observations	=COUNTIF(DATA!A2:A63,"Female")	33	The Female group contains 33 records.
Male observations	=COUNTIF(DATA!A2:A63,"Male")	29	The Male group contains 29 records.

Why COUNT and COUNTA are different

COUNTA counts any non-blank entry, including text. COUNT counts only numeric values. Therefore, if COUNTA reported 62 WAM entries but COUNT reported fewer than 62, one or more WAMs might be blank or stored as text.

2.2 Inspect the Gender entries

1. **Return to DATA and click cell A1.** Choose Data > Filter. A drop-down arrow appears in the Gender header.
2. **Open the Gender filter.** The list should contain only Female and Male. Check for blank entries, different capitalisation, extra spaces, or misspellings such as “female”, “Male ” or “Fmale”.
3. **Clear the filter after checking.** Choose Clear Filter From “Gender” so all rows are visible again.

Final finding: the supplied dataset contains exactly two recorded categories—Female and Male—with no blank Gender cells and no unexpected spelling visible in the source data.

2.3 Inspect the WAM entries

1. **Click any WAM cell in column F.** A real number normally aligns to the right under Excel’s default settings. Text often aligns to the left and may display a green warning triangle.
2. **Check the possible range.** In a spare area of Part 3 Analysis, type =MIN(DATA!F2:F63) and =MAX(DATA!F2:F63). The expected results are 27.9 and 93.9. Both are within the possible WAM range of 0 to 100.
3. **Check impossible values directly.** Type =COUNTIF(DATA!F2:F63,"<0") and =COUNTIF(DATA!F2:F63,">100"). Both should return 0.

```
=MIN(DATA!F2:F63)   →   27.9
=MAX(DATA!F2:F63)   →   93.9
=COUNTIF(DATA!F2:F63,"<0")   →   0
=COUNTIF(DATA!F2:F63,">100")   →   0
```

2.4 What to do with an unusual value

A WAM of 27.9 is low compared with most observations, and 93.9 is high. This does not prove either value is wrong. Before changing or deleting an observation, verify:

- the value against the original academic record or source questionnaire;
- whether WAM was defined consistently for every respondent;
- whether a decimal point or digit was entered incorrectly;
- whether the value belongs to the correct respondent and row; and
- whether the observation is genuinely extreme but valid.

Checkpoint before moving on

Audit results: 62 non-blank Gender entries; 62 numeric WAMs; 0 missing WAMs; 33 Female and 29 Male records; WAM minimum 27.9 and maximum 93.9; no WAM below 0 or above 100.

Task 3 — Create the Sorted Data worksheet

The data must be sorted by Gender so that the Female WAMs and Male WAMs form two continuous input ranges for the Analysis ToolPak.

1. **Select the complete source block.** On DATA, click A1. Hold Shift and click N63, or type A1:N63 into the Name Box to the left of the Formula Bar and press Enter.
2. **Copy the full block.** Press Ctrl+C on Windows or Command+C on Mac.
3. **Paste into Sorted Data.** Click the Sorted Data tab, click A1 and press Ctrl+V or Command+V.
4. **Select the complete pasted block.** Select A1:N63. Every column must be included, not only Gender.
5. **Open the Sort dialog.** Choose Data > Sort. Tick My data has headers.
6. **Specify the sort.** Sort by Gender; Sort On Cell Values; Order A to Z. Click OK.
7. **Confirm the final row ranges.** Female alphabetically precedes Male, so Female records occupy rows 2–34 and Male records occupy rows 35–63. WAM is column F.

Item	Confirmed range	Check
Female WAM	'Sorted Data'!F2:F34	33 observations
Male WAM	'Sorted Data'!F35:F63	29 observations
WAM column	F	Use the same column for both groups

Common mistake to avoid

Never sort only column A. If Gender moves but WAM, Age, Major and the other columns stay in their original rows, each respondent's attributes become attached to the wrong person. That corrupts the dataset even though the worksheet may still look tidy.

Checkpoint before moving on

Click several rows after sorting and confirm that every row moved as one complete record. Female is in A2:A34; Male is in A35:A63; Female WAM is F2:F34; Male WAM is F35:F63.

Task 4 — Enable the Analysis ToolPak

Windows instructions

1. **Choose File > Options.**
2. **Select Add-ins on the left.**
3. **At the bottom, set Manage to Excel Add-ins and click Go.**
4. **Tick Analysis ToolPak and click OK.**
5. **Return to the Data tab. Data Analysis should appear near the right-hand side.**

Mac instructions

1. **Choose Tools > Excel Add-ins.**
2. **Tick Analysis ToolPak.**
3. **Click OK.**
4. **Return to the Data tab and confirm that Data Analysis is visible.**

Troubleshooting

If Data Analysis still does not appear, close and reopen Excel. Check that you are using desktop Excel rather than Excel Online. On a managed university computer, you may need to sign in or ask IT support to install the add-in.

Task 5 — Produce descriptive statistics for Female and Male WAM

Descriptive statistics summarise the centre, variability and shape of a distribution. You must run the ToolPak twice—once for Female WAM and once for Male WAM—using the exact continuous ranges created by the sort.

5.1 Run the Female descriptive statistics

1. **Open the ToolPak.** Click Data > Data Analysis.
2. **Select Descriptive Statistics.** Click Descriptive Statistics, then click OK.
3. **Enter the Female Input Range.** Click in Input Range, switch to Sorted Data and select F2:F34. Excel should display 'Sorted Data'!$F$2:$F$34.
4. **Choose Grouped By: Columns.** The WAM values are arranged vertically in one column.
5. **Leave Labels in first row unticked.** Cell F2 is a number, not the word WAM.
6. **Choose an Output Range.** Switch to Part 3 Analysis and choose a blank starting cell. In the supplied completed workbook, the clean comparison table is in A18:E34 and ToolPak-style outputs are also shown to the far right.
7. **Tick Summary statistics.** This requests mean, median, standard deviation and the other required measures.
8. **Tick Confidence Level for Mean and enter 95%.** Excel reports the confidence amount, not the lower and upper endpoints.
9. **Click OK.** Check that Count equals 33. If it does not, the input range is wrong.

5.2 Run the Male descriptive statistics

1. **Open Data Analysis again.** Choose Data > Data Analysis > Descriptive Statistics.
2. **Enter the Male Input Range.** Use 'Sorted Data'!$F$35:$F$63.
3. **Use the same settings.** Grouped By Columns; Labels unticked; Summary statistics ticked; Confidence Level for Mean 95%.
4. **Choose a different blank output area.** Do not overwrite the Female output.
5. **Click OK and check Count.** The Male Count must equal 29.

5.3 Final descriptive-statistics results

Measure	Female	Male	Plain-language interpretation
Mean	64.715	63.059	Average WAM; Female is 1.657 marks higher.
Standard Error	2.388	2.221	Estimated sampling variability of each sample mean.
Median	64.9	62.4	Middle WAM; Female is 2.5 marks higher.
Mode(s)	No mode	56.3 and 72.5	Female has no repeated WAM; Male is bimodal.
Standard Deviation	13.717	11.960	Typical spread around the mean; Female is more dispersed.

Measure	Female	Male	Plain-language interpretation
Sample Variance	188.156	143.053	Standard deviation squared.
Kurtosis	0.885	-0.439	Female has heavier tails relative to a normal distribution.
Skewness	-0.177	-0.074	Both are close to zero; neither is strongly skewed.
Range	66.0	48.0	Maximum minus minimum.
Minimum	27.9	38.9	Low values that merit verification.
Maximum	93.9	86.9	High values that may also be verified.
Sum	2135.6	1828.7	Total of all WAM values in each group.
Count	33	29	Must agree with Task 2.
Confidence Level (95.0%)	4.864	4.550	Margin added to and subtracted from the mean.
95% CI lower	59.851	58.509	Mean minus confidence amount.
95% CI upper	69.579	67.608	Mean plus confidence amount.

5.4 Required Female-minus-Male calculations

Calculation	Excel logic	Final result
Mean difference	= Female mean − Male mean	64.715 − 63.059 = 1.657
Median difference	= Female median − Male median	64.9 − 62.4 = 2.5
Standard-deviation difference	= Female SD − Male SD	13.717 − 11.960 = 1.757
Female CI lower	= Female mean − Female confidence amount	64.715 − 4.864 = 59.851
Female CI upper	= Female mean + Female confidence amount	64.715 + 4.864 = 69.579
Male CI lower	= Male mean − Male confidence amount	63.059 − 4.550 = 58.509
Male CI upper	= Male mean + Male confidence amount	63.059 + 4.550 = 67.608

5.5 Mode check: do not rely only on ToolPak

A mode is a value that occurs more often than other values. Excel's ToolPak may print only one value when several values are tied for the highest frequency. Therefore, inspect the raw data or use a frequency count.

- Female: every WAM occurs once, so there is no mode.
- Male: 56.3 occurs twice and 72.5 occurs twice. Therefore, the Male distribution has two modes.
- The ToolPak may display only 56.3. Reporting only that value would be incomplete.

5.6 What the confidence amount means

The ToolPak line "Confidence Level (95.0%)" is a margin, not a complete interval. Construct endpoints by subtracting and adding the margin to the mean.

```
Female: 64.715 ± 4.864 = [59.851, 69.579]
Male: 63.059 ± 4.550 = [58.509, 67.608]
```

Interpret cautiously

A confidence interval has its usual population-inference interpretation only when the sampling design is defensible and the assumptions are appropriate. If respondents form a convenience or self-selected sample, the calculation can still be reported as a numerical summary, but generalisation beyond this sample is limited.

Checkpoint before moving on

Female Count = 33 and Male Count = 29. Mean difference = 1.657, median difference = 2.5 and standard-deviation difference = 1.757. Female CI = [59.851, 69.579]; Male CI = [58.509, 67.608].

Task 6 — Construct identical WAM bins and a common-bin frequency table

A histogram groups observations into intervals. The same bin boundaries must be used for Female and Male so that bar heights represent directly comparable intervals.

6.1 Enter the bin upper limits

1. **Choose a blank table area on Part 3 Analysis.** In the completed workbook, the common-bin table starts at A38.
2. **Type the bin upper limits.** Enter 10 in A39 and 20 in A40. Select both cells, drag the fill handle downward until Excel reaches 100.
3. **Create readable interval labels.** In the next column, type the labels exactly as shown below. Remember that Excel treats the bin number as an upper limit.

Bin upper limit	WAM interval	Female frequency	Male frequency
10	WAM ≤ 10	0	0
20	10 < WAM ≤ 20	0	0
30	20 < WAM ≤ 30	1	0
40	30 < WAM ≤ 40	0	1
50	40 < WAM ≤ 50	2	3
60	50 < WAM ≤ 60	9	8
70	60 < WAM ≤ 70	11	7
80	70 < WAM ≤ 80	5	7
90	80 < WAM ≤ 90	4	3
100	90 < WAM ≤ 100	1	0
Total		**33**	**29**

6.2 Run the Female Histogram tool

1. **Choose Data > Data Analysis > Histogram.**
2. **Input Range** Select 'Sorted Data'!$F$2:$F$34.
3. **Bin Range** Select the cells containing 10, 20, 30, …, 100.
4. **Labels** Leave unticked because the selected ranges contain numbers only.
5. **Output Range** Choose a blank starting cell on Part 3 Analysis.
6. **Chart Output** You may leave this unticked because Task 7 creates cleaner charts from the common table.
7. **Click OK** The ten bin frequencies must sum to 33. A “More” row, if displayed, should equal 0 because all WAMs are at most 100.

6.3 Repeat for Male

Repeat the Histogram procedure using Input Range 'Sorted Data'!$F$35:$F$63 and exactly the same Bin Range. The Male frequencies must sum to 29. Do not change bin widths or upper limits.

6.4 Optional formula cross-check

The ToolPak is required for practice, but formulas provide a transparent cross-check. For example:

Female WAM ≤ 10: `=COUNTIF('Sorted Data'!$F$2:$F$34,"<=10")`

Female 10 < WAM ≤ 20: `=COUNTIFS('Sorted Data'!$F$2:$F$34,">10",'Sorted Data'!$F$2:$F$34,"<=20")`

Change the lower and upper limits for later bins. Use the Male range F35:F63 for the Male column.

Checkpoint before moving on

Female frequencies total 33 and Male frequencies total 29. These totals agree with the dataset audit and descriptive-statistics Count values.

Common mistake to avoid

Do not use one set of bins for Female and a different set for Male. Do not compare histograms whose axes use different maximum values. Either mistake can make one distribution appear more or less variable simply because of chart design.

Task 7 — Create the final three figures

7.1 Combined Female–Male chart

1. **Select the chart data.** Select the readable interval labels and both frequency columns, excluding the Total row. In the completed workbook this is B38:D48 if row 38 contains headers, or B39:D48 if you select data only.
2. **Insert the chart.** Choose Insert > Insert Column or Bar Chart > 2-D Column > Clustered Column.
3. **Set the title.** Click the chart title and type WAM distributions using identical 10-mark bins.
4. **Add axis titles.** Choose Chart Design > Add Chart Element > Axis Titles. Horizontal: WAM interval. Vertical: Number of students.
5. **Keep the legend.** The legend must identify Female and Male.
6. **Set the vertical-axis scale.** Right-click the vertical axis > Format Axis. Under Axis Options set Minimum = 0, Maximum = 12 and Major unit = 2. Use Fixed values, not Auto.
7. **Move and resize carefully.** Place the chart where it does not cover the audit, statistics or frequency tables.

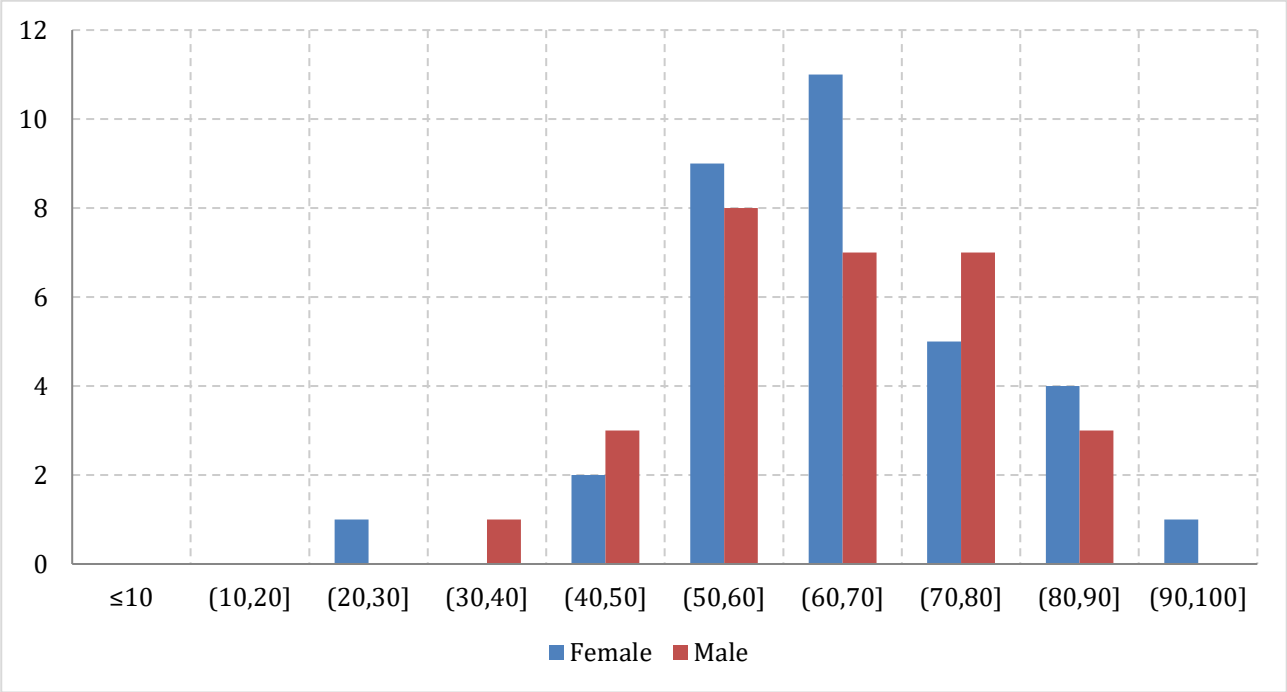

Figure 1. Combined Female and Male WAM frequencies using identical 10-mark bins and a common 0–12 vertical scale

7.2 Separate Female chart

Select the interval labels and Female frequency column only. Insert a Clustered Column chart. Use the title Female WAM distribution (n = 33). Use exactly the same horizontal labels and vertical-axis settings: Minimum 0, Maximum 12, Major unit 2.

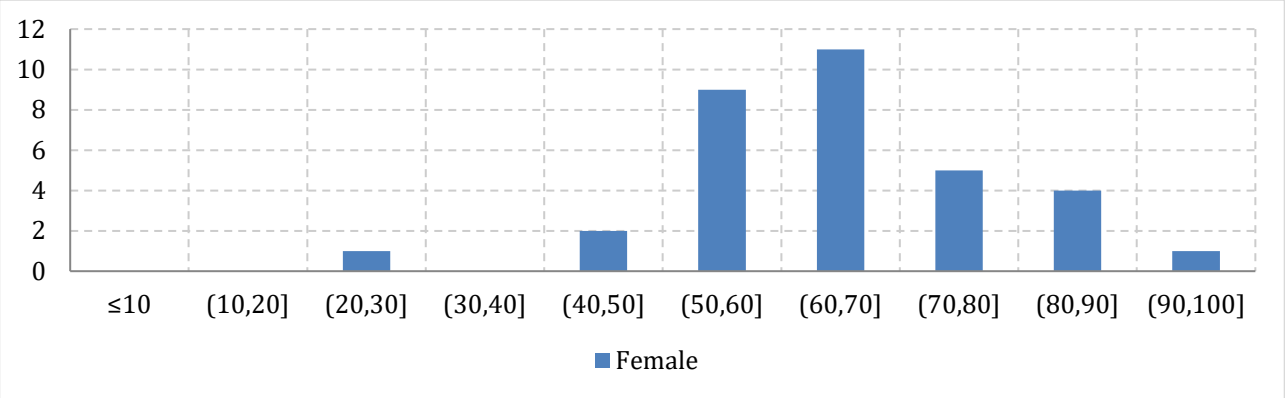

Figure 2. Female WAM distribution (n = 33)

7.3 Separate Male chart

The easiest method is to copy the Female chart, paste it, then choose Chart Design > Select Data. Edit the series so the values refer to the Male frequency column. Change the title to Male WAM distribution (n = 29). Keep Minimum 0, Maximum 12 and Major unit 2.

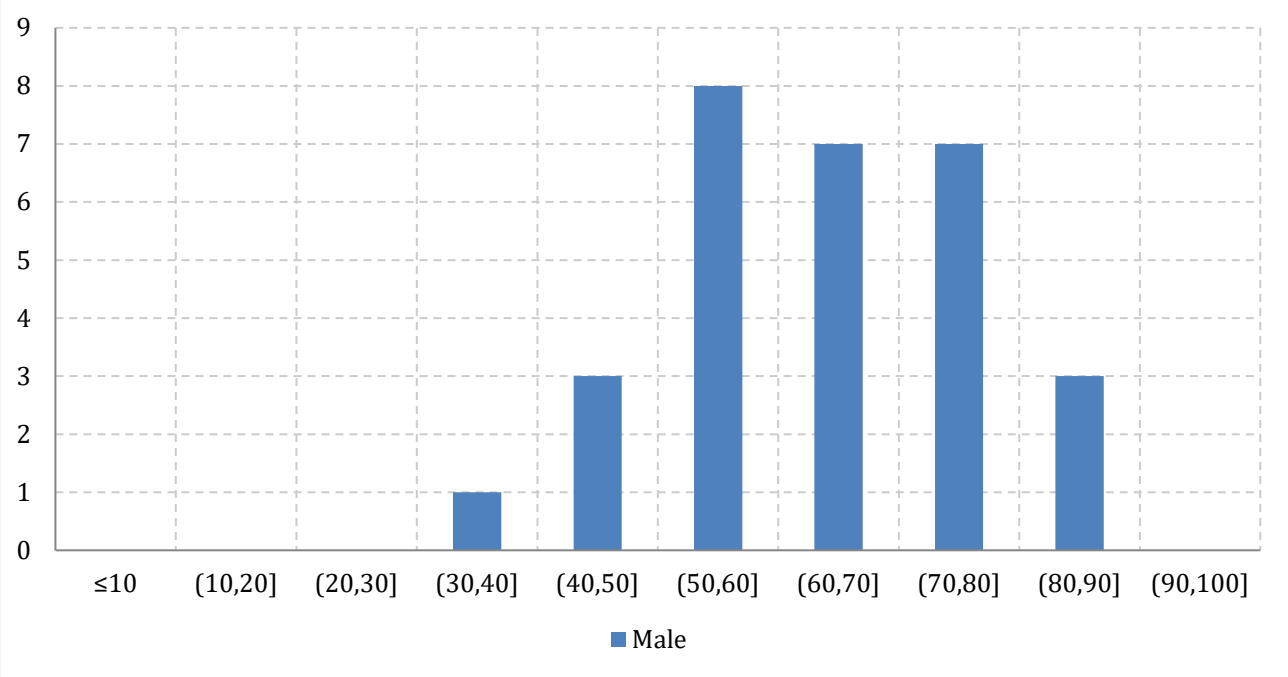

Figure 3. Male WAM distribution (n = 29)

7.4 Why identical bins and scales matter

A chart is a visual measurement device. Changing the bin width changes the grouping of observations; changing the vertical-axis maximum changes the apparent height of bars. If the Male chart stopped at 8 but the Female chart stopped at 12, the Male bars would look artificially tall even when their frequencies were lower. Identical bins and identical scales prevent this distortion.

Checkpoint before moving on

Three charts are present: one combined chart, one Female chart and one Male chart. Every chart uses the ten required intervals and a vertical axis from 0 to 12 in steps of 2.

Task 8 — Interpret the final tables and figures

Use complete sentences and cite the actual numbers from your workbook. The model answers below are written at an appropriate level for submission.

(a) Dataset composition

There are 62 usable numeric WAM observations and 0 missing WAM values. The recorded-gender groups contain 33 Female respondents and 29 Male respondents. The difference is only four observations, so the sample is reasonably balanced for a descriptive comparison, although it is not perfectly equal.

(b) Centre

The Female mean WAM is 64.715 and the Male mean is 63.059, giving Female minus Male = 1.657 marks. The Female median is 64.9 and the Male median is 62.4, giving a difference of 2.5 marks. These differences are small relative to the within-group standard deviations of 13.717 and 11.960 marks.

(c) Variability

The Female group is more dispersed. Its sample standard deviation is 13.717 compared with 11.960 for Male. Its range is also wider: 66.0 marks, from 27.9 to 93.9, compared with 48.0 marks for Male, from 38.9 to 86.9.

(d) Shape

Neither distribution is strongly skewed. Female skewness is -0.177 and Male skewness is -0.074, both close to zero. The histograms show most observations concentrated between 50 and 80, with some observations in both tails. Female kurtosis is 0.885, suggesting somewhat heavier tails than a normal distribution, whereas Male kurtosis is -0.439.

(e) Unusual observations

The Female WAM of 27.9 and maximum of 93.9 merit checking; the Male minimum of 38.9 and maximum of 86.9 may also be verified. Before changing a value, confirm it against the original academic record, the WAM definition and the data-entry process. An unusual value may be genuine and should not be deleted merely because it is extreme.

(f) Modes

The Female group has no mode because every WAM value occurs once. The Male group has two modes: 56.3 and 72.5, each occurring twice. The Analysis ToolPak may display only one tied mode, so the raw values or a frequency count must be inspected before concluding that a distribution has only one mode.

(g) Frequency-table reading

The most common Female interval is 60 < WAM ≤ 70 with 11 observations. The most common Male interval is 50 < WAM ≤ 60 with 8 observations. In 70 < WAM ≤ 80 there are 5 Female and 7 Male observations. Intervals containing exactly one observation are 20 < WAM ≤ 30 and 90 < WAM ≤ 100 for Female, and 30 < WAM ≤ 40 for Male.

(h) Confidence intervals

Using the ToolPak confidence amounts, the Female 95% interval is 59.851 to 69.579 and the Male interval is 58.509 to 67.608. These intervals overlap substantially. They should be interpreted cautiously because valid population inference requires an appropriate probability sample, independent observations and a defensible sampling frame. A convenience or self-selected sample cannot support unrestricted generalisation.

(i) Overall conclusion

In this sample, Female respondents have a slightly higher mean and median WAM, while their WAM values are also somewhat more dispersed. The centre differences are small compared with variation within each group. These data do not establish that recorded gender causes academic-performance differences. Potential confounders include major, year of study, age, enrolment status, employment hours, prior preparation and other unmeasured factors. The analysis is descriptive, not causal.

Task 9 — Final submission and error checks

Submit one Excel workbook, not separate screenshots. The workbook must preserve the calculations that produced the figures.

Check	Requirement
1	DATA worksheet remains unchanged.
2	Sorted Data contains the complete A1:N63 block sorted by Gender using full rows.
3	Part 3 Analysis contains the five required audit formulas and results.
4	The descriptive-statistics comparison includes all required ToolPak measures.
5	Female-minus-Male mean, median and standard-deviation differences are calculated.
6	The confidence amounts are converted into lower and upper endpoints.
7	The frequency table uses common upper limits 10, 20, …, 100.
8	Female frequencies sum to 33 and Male frequencies sum to 29.
9	One combined and two separate charts use identical bins and vertical scale 0–12.
10	All nine Task 8 interpretation responses are written in complete sentences.

Final reconciliation rule
The Female descriptive-statistics Count, Female frequency total and Female audit count must all equal 33. The corresponding three Male values must all equal 29. Any mismatch means that an input range, row or bin range was selected incorrectly.

Appendix — Excel formulas used in the completed comparison table

The tutorial requires the Analysis ToolPak. The formulas below are included as a reproducibility check and help explain how each statistic is constructed. Dollar signs make the ranges absolute so they do not move when formulas are copied.

Measure	Female formula	Male formula
Mean	`=AVERAGE('Sorted Data'!$F$2:$F$34)`	`=AVERAGE('Sorted Data'!$F$35:$F$63)`
Standard Error	`=STDEV.S('Sorted Data'!$F$2:$F$34)/SQRT(COUNT('Sorted Data'!$F$2:$F$34))`	`=STDEV.S('Sorted Data'!$F$35:$F$63)/SQRT(COUNT('Sorted Data'!$F$35:$F$63))`
Median	`=MEDIAN('Sorted Data'!$F$2:$F$34)`	`=MEDIAN('Sorted Data'!$F$35:$F$63)`
Standard deviation	`=STDEV.S('Sorted Data'!$F$2:$F$34)`	`=STDEV.S('Sorted Data'!$F$35:$F$63)`
Sample variance	`=VAR.S('Sorted Data'!$F$2:$F$34)`	`=VAR.S('Sorted Data'!$F$35:$F$63)`
Kurtosis	`=KURT('Sorted Data'!$F$2:$F$34)`	`=KURT('Sorted Data'!$F$35:$F$63)`
Skewness	`=SKEW('Sorted Data'!$F$2:$F$34)`	`=SKEW('Sorted Data'!$F$35:$F$63)`
Range	`=MAX('Sorted Data'!$F$2:$F$34)-MIN('Sorted Data'!$F$2:$F$34)`	`=MAX('Sorted Data'!$F$35:$F$63)-MIN('Sorted Data'!$F$35:$F$63)`
Minimum	`=MIN('Sorted Data'!$F$2:$F$34)`	`=MIN('Sorted Data'!$F$35:$F$63)`
Maximum	`=MAX('Sorted Data'!$F$2:$F$34)`	`=MAX('Sorted Data'!$F$35:$F$63)`
Sum	`=SUM('Sorted Data'!$F$2:$F$34)`	`=SUM('Sorted Data'!$F$35:$F$63)`
Count	`=COUNT('Sorted Data'!$F$2:$F$34)`	`=COUNT('Sorted Data'!$F$35:$F$63)`
95% confidence amount	`=CONFIDENCE.T(0.05,STDEV.S('Sorted Data'!$F$2:$F$34),COUNT('Sorted Data'!$F$2:$F$34))`	`=CONFIDENCE.T(0.05,STDEV.S('Sorted Data'!$F$35:$F$63),COUNT('Sorted Data'!$F$35:$F$63))`

Completed-workbook layout

Worksheet	Range	Content
DATA	A1:N63	Original source, unchanged
Sorted Data	A1:N63	Complete rows sorted Female then Male
Sorted Data	P1:R5	Confirmed row-range notes
Part 3 Analysis	A3:I9	Dataset audit and additional WAM checks
Part 3 Analysis	A17:J34	Clean descriptive-statistics table, differences and confidence intervals
Part 3 Analysis	A37:J49	Common-bin frequency table and validation notes
Part 3 Analysis	N2:U57	Combined, Female and Male charts
Part 3 Analysis	A52:U61	Complete Task 8 model answers
Part 3 Analysis	W2:AA17	ToolPak-style outputs for line-by-line comparison

Final descriptive conclusion

The Female sample has a mean WAM about 1.657 marks higher and a median 2.5 marks higher than the Male sample. Female WAM is also more variable. Because the differences in centre are small relative to standard deviations of roughly 12–14 marks, and because this is an observational sample, the correct conclusion is descriptive rather than causal.