

STATISTICAL LABORATORY

Lecture 5: Measures of Skewness (Theory and Practice)

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

`allenlui@iseg.ulisboa.pt`

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

- 1 Fundamental Concepts of Statistics
- 2 Exploratory Data Analysis
- 3 Organizing and Summarizing Data**
- 4 Association and Relationships Between Variables
- 5 Index Numbers
- 6 Time Series Analysis

Exercise 3.2

Consider the information provided in exercise 2.3

Table: Monthly wages of workers in a textile company

Salary (euros)	Workers (%)
[600; 900]	9
]900; 1000]	18
]1000; 1100]	35
]1100; 1200]	22
]1200; 1300]	16

- a) Assess the symmetry (or asymmetry) of the distribution using a boxplot.
- b) Determine the new mean salary after a general wage increase of 10%.
- c) Evaluate the concentration of the data.

Exercise 3.4

The information on the salaries of workers in a multinational company is presented below:

Classes	0 – 3	3 – 6	6 – 9	9 – 12
Number of Workers	500	1500	2500	500

- a) Compute the mean and median salaries.
- b) What percentage of workers earn more than the mean salary?
- c) Determine the mean salary of the 30% of workers with the lowest salaries.
- d) Determine the mean salary of the 30% of workers with the highest salaries.

Exercise 3.5

The statistical information on the grades of a group of university students in Statistics and Economics is presented below:

Student	A	B	C	D	E	F	G	H	I	J
Statistics	10	15	13	16	15	12	11	14	10	11
Economics	12	13	13	15	16	13	12	18	12	15

- a) Assess the variability of grades in each subject individually.
- b) Assess the variability of grades in these subjects comparatively.

Exercise 3.6

The data below show the diesel consumption (litres) of 10 cars of brands A and B over a 100 km trip:

Brand A: 7,8; 7,0; 8,2; 7,6; 6,9; 7,7; 7,2; 7,8; 7,3; 7,5

Brand B: 7,0; 7,1; 7,1; 7,3; 7,2; 7,0; 7,8; 8,0; 7,0; 7,0

- a) Compare the average consumption using an appropriate measure of central tendency.
- b) Using a back-to-back stem-and-leaf plot, assess the symmetry of the data and interpret it in context.
- c) Using the cumulative frequency function, compute the proportion of brand A cars whose consumption exceeds 7.3 litres. Identify the corresponding frequency.

Skewness

- Skewness measures the asymmetry of a probability distribution.
- Identifying the presence and direction of skewness is essential across statistics for several distinct reasons:
- Impact on Central Tendency and Summaries:
 - Mean-Median Divergence: In skewed distributions, the mean is pulled into the long tail, whereas the median remains a robust indicator of central location.
 - Misleading Averages: Relying solely on the mean for heavily skewed data (such as income distributions or firm sizes) distorts the perception of a "typical" observation.

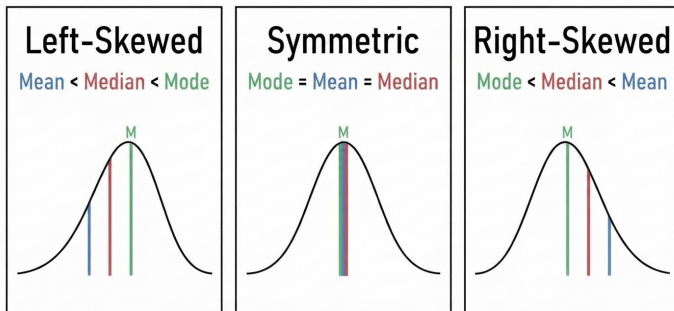
Consequences for Statistical Inference:

- Symmetry Violations: Many standard estimators, test statistics, and confidence intervals implicitly assume symmetric structures.
- Biased Inference: Unaccounted skewness can invalidate statistical inference, as many inferential method are based on normal distribution (which is a symmetric distribution) assumption.

Identifying skewness using mean, median, and mode

Skewness can be identified by the relationship between Mode, Median, and Mean.

- $\text{Mean} > \text{Median} > \text{Mode} \rightarrow$ positively skewed (right-skewed)
- $\text{Mean} < \text{Median} < \text{Mode} \rightarrow$ negatively skewed (left-skewed)
- $\text{Mean} \approx \text{Median} \approx \text{Mode} \rightarrow$ approximately symmetric



Pearson's coefficient of skewness

According to **Silvestre (2007)**, the coefficient of skewness is defined as:

$$g = \frac{\bar{x} - Mo}{s}$$

where:

- $\bar{x}$ = sample mean
- Mo = mode
- s = standard deviation

Interpretation:

- $g = 0$: symmetric distribution
- $g > 0$: positively skewed (right tail)
- $g < 0$: negatively skewed (left tail)

Pearson's coefficient of skewness

- Pearson's coefficient measures how far the mean ($\bar{x}$) is pulled away from the peak or mode (Mo), standardized by the standard deviation (s).
- In a symmetrical distribution, the mean and mode coincide ($g = 0$).
- In a positively skewed distribution, extreme high values pull the mean above the mode, making the numerator positive ($g > 0$).
- The formula implicitly assumes a unimodal distribution.

Bowley's Coefficient of Skewness

According to Silvestre (2007), Bowley's coefficient of skewness is defined as:

$$g' = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

where:

- Q_1 = first quartile
- Q_2 = median (second quartile)
- Q_3 = third quartile

Interpretation:

- $g' = 0$: symmetric distribution
- $g' > 0$: positively skewed (right tail)
- $g' < 0$: negatively skewed (left tail)

Note

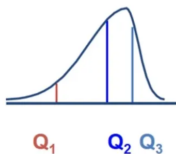
This measure is based on quartiles, and therefore less affected by extreme values compared to Pearson's coefficient.

Bowley's Coefficient of Skewness

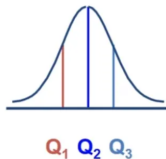
- Bowley's coefficient looks at quartiles (Q_1, Q_2, Q_3) to compare the spread of the upper half of the data against the lower half.
- It measures the distance from the median to the third quartile ($(Q_3 - Q_2)$) versus the median to the first quartile ($(Q_2 - Q_1)$).
- If the upper half is stretched out wider than the lower half, the numerator becomes positive, indicating a right tail.

Bowley's Coefficient of Skewness

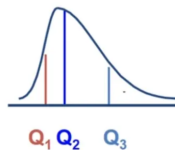
Left-Skewed



Symmetric



Right-Skewed



Moment coefficient of skewness

According to Silvestre (2007), the moment coefficient of skewness is defined as:

$$b_1 = \frac{m_3^2}{m_2^3}$$

or equivalently,

$$g_1 = \sqrt{b_1} = \frac{m_3}{m_2^{3/2}}$$

where:

- m_2 = second central moment about the mean (variance)
- m_3 = third central moment about the mean, given by:

Moment coefficient of skewness (cont.)

- $m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$ (for ungrouped data)
- $m_3 = \frac{1}{n} \sum_{j=1}^m n_j (x'_j - \bar{x})^3$ (for grouped data)

where:

- m = number of classes
- x'_j = class midpoint
- n_j = class frequency
- $\bar{x}$ = grouped mean

Interpretation:

- $g_1 = 0$: symmetric distribution
- $g_1 > 0$: positively skewed (right tail)
- $g_1 < 0$: negatively skewed (left tail)

Moment coefficient of skewness (cont.)

- This measure relies on central moments, specifically looking at cubed deviations from the mean $(x_i - \bar{x})^3$.
- Cubing is crucial because it preserves the mathematical sign (values above the mean stay positive, values below stay negative) while heavily weighting values that are farther away from the center.
- Summing these cubed deviations (m_3) captures the net imbalance of the tails, which is then scaled relative to the distribution's dispersion ($m_2^{3/2}$).

Comparison of skewness coefficients

Coefficient	Formula	Based on	Advantages	Disadvantages
Pearson's g	$g = \frac{\bar{x} - Mo}{s}$	Mean, mode, standard deviation	Easy to compute; intuitive	Sensitive to outliers; requires mode estimation
Bowley's g'	$g' = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$	Quartiles (Q_1 , Q_2 , Q_3)	Robust to extreme values; useful for ordinal data	Less precise for small datasets; ignores all data outside quartiles
Moment g_1	$g_1 = \frac{m_3}{m_2^{3/2}}$	Central moments about the mean	Theoretical; widely used in inferential statistics	Sensitive to outliers; requires all data points

Interpretation (for all):

- $g = 0$: symmetric distribution
- $g > 0$: positively skewed (right tail)
- $g < 0$: negatively skewed (left tail)

Summary: identifying skewness

- Histogram → continuous variables
- Bar chart → discrete or ordinal variables (only attempt to measure skewness graphically using bar charts if the scale has at least 5 distinct ordered categories.)
- Boxplot → observe the **position of the median** relative to Q1 and Q3
- Another way to identify skewness is through the relationship between **mean, median, and mode**:
 - Symmetric: $\text{mean} \approx \text{median} \approx \text{mode}$
 - Positively skewed: $\text{mean} > \text{median} > \text{mode}$
 - Negatively skewed: $\text{mean} < \text{median} < \text{mode}$

Summary: identifying skewness (Continued)

- **Pearson's coefficient (g)**: mean and mode (requires numeric coding)
- **Bowley's coefficient (g')**: quartiles (robust, suitable for ordinal data)
- **Moment coefficient (g_1)**: central moments (requires numeric coding)
- Graphs give a **visual impression** of skewness
- Measures and coefficients give a **quantitative assessment**
- For ordinal variables with 5+ classes, **boxplots** and **Bowley's coefficient** are particularly useful

Example I: study of skewness for ungrouped data

Data (number of complaints in 10 firms):

$$X = [2, 3, 3, 4, 4, 5, 5, 6, 7, 9]$$

Step 1: Basic statistics

- Mean: $\bar{x} = 4.8$
- Median: $Me = 4.5$
- Mode: $Mo = 4$

Step 2: Skewness by relation of measures

- Mean $>$ Median $>$ Mode $\rightarrow$ **positively skewed**

Step 3: Pearson's coefficient

- Standard deviation: $s \approx 1.87$

$$g = \frac{\bar{x} - Mo}{s} = \frac{4.8 - 4}{1.87} \approx 0.43$$

Example I: study of skewness for ungrouped data (Continued)

Step 4: Bowley's coefficient (quartiles)

- $Q1 = 3.25, Q2 = 4.5, Q3 = 6$

$$\begin{aligned} g' &= \frac{(Q3 - Q2) - (Q2 - Q1)}{(Q3 - Q2) + (Q2 - Q1)} = \frac{(6 - 4.5) - (4.5 - 3.25)}{(6 - 4.5) + (4.5 - 3.25)} \\ &= \frac{1.5 - 1.25}{1.5 + 1.25} = \frac{0.25}{2.75} \approx 0.09. \end{aligned} \quad (1)$$

Step 5: Moment coefficient

- Second central moment: $m_2 = s^2 = 3.5$
- Third central moment: $m_3 \approx 0.53$

$$g_1 = \frac{m_3}{m_2^{3/2}} \approx \frac{0.53}{3.5^{1.5}} \approx 0.09$$

Conclusion:

- All three methods indicate a **slight positive skewness**, consistent with **mean > median > mode**.

Example 2: study of skewness for grouped data

Example: Skewness Analysis for Grouped Data (3 Classes)

Data

Class interval	Midpoint x'_j	Freq. n_j	Relative freq. f_j	Cumulative rel. freq. F_j^*
[0–10]	5	6	0.200	0.200
(10–20]	15	12	0.400	0.600
(20–30]	25	12	0.400	1.000

Total $n = 30$, $h = 10$.

Example 2: study of skewness for grouped data (cont.)

1) Mean

$$\bar{X} = \sum f_j x'_j = 0.2 \cdot 5 + 0.4 \cdot 15 + 0.4 \cdot 25 = 17.0$$

2) Median

Median class: first class where $F_j^* \geq 0.5 \rightarrow (10-20]$

$$\text{Med} = l + \frac{0.5 - F^*(l)}{F^*(L) - F^*(l)} h = 10 + \frac{0.5 - 0.2}{0.6 - 0.2} \cdot 10 = 10 + 7.5 = 17.5$$

3) Mode

Modal class: class with largest $f(L) \rightarrow (10-20]$

$$f^* = f(L) - f(l) = 0.4 - 0.2 = 0.2, \quad f^{**} = f(L) - f(L+1) = 0.4 - 0.4 = 0$$

$$\text{Mo} = l + \frac{f^*}{f^* + f^{**}} h = 10 + \frac{0.2}{0.2 + 0} \cdot 10 = 20.0$$

Example 2: study of skewness for grouped data (cont.)

4) Skewness Coefficients

- Pearson: $g = \frac{\bar{X} - Mo}{s}$
 - Variance $m_2 = \sum f_j(x'_j - \bar{X})^2 = 44.0$
 - SD $s = \sqrt{m_2} \approx 6.633$
 - $g \approx \frac{17.0 - 20.0}{6.633} \approx -0.452$ (negative skewness)
- Bowley (quartile-based):
 - $Q_1 \approx 11.667$, $Q_2 = \text{Med} = 17.5$, $Q_3 \approx 23.333$
 - $g' = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{(23.333 - 17.5) - (17.5 - 11.667)}{(23.333 - 17.5) + (17.5 - 11.667)} = 0$ (symmetric by quartiles)
- Moment coefficient:

$$g_1 = \frac{m_3}{m_2^{3/2}}, \quad m_3 = \sum f_j(x'_j - \bar{X})^3 = 0$$
$$\Rightarrow g_1 = 0 \text{ (moment indicates symmetry)}$$

Example 2: study of skewness for grouped data (cont.)

Conclusion:

- $\text{Mean} < \text{Median} < \text{Mode} \rightarrow$ suggests slight negative skew
- Bowley and moment coefficients $= 0 \rightarrow$ quartiles and central moments indicate approximate symmetry
- Highlights how different skewness measures can give slightly different indications for small sample/grouped data.

Exercise I: ungrouped data

Exercise 1: Ungrouped Data

The number of complaints in 12 firms is recorded as:

$$X = [1, 2, 2, 3, 4, 4, 5, 5, 6, 7, 8, 9]$$

Tasks:

- 1 Compute the mean, median, and mode.
- 2 Determine the skewness direction using the relationship mean–median–mode.
- 3 Calculate the Pearson's coefficient of skewness.
- 4 Calculate the Bowley's coefficient (use quartiles).
- 5 Calculate the moment coefficient of skewness.
- 6 Compare the results and interpret the overall skewness.

EXERCISE I: SOLUTION

Answer:

Example 1 — Ungrouped data

Data: $X = [1, 2, 2, 3, 4, 4, 4, 5, 5, 6, 7, 8, 9]$ ($n = 12$)

Basic stats

- $\bar{x} = 4.667$
- Median = 4.5
- Mode = 4 (tie: 2, 4, 5 — we choose 4)
- Second central moment (variance, population)
 $m_2 = 5.722 \rightarrow s = \sqrt{m_2} = 2.392$
- Third central moment $m_3 = 3.426$

Relation mean–median–mode

$\bar{x} > \text{median} > \text{mode} \rightarrow \text{positively skewed (right skew)}$

Exercise 1: Solution

Pearson's coefficient

$$g = \frac{\bar{x} - Mo}{s} = \frac{4.667 - 4}{2.392} \approx 0.28$$

Bowley's coefficient (quartile-based) Quartiles (by standard method):

$$Q_1 = 2.5, Q_2 = 4.5, Q_3 = 6.5$$

$$g' = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{(6.5 - 4.5) - (4.5 - 2.5)}{(6.5 - 4.5) + (4.5 - 2.5)} = 0$$

(Here Bowley = 0 — quartiles are symmetric even though mean/median/mode and moments show slight positive skew. Bowley can be zero for small samples / particular quartile positions.)

Exercise 1: Solution (continued)

Moment coefficient

$$g_1 = \frac{m_3}{m_2^{3/2}} \approx \frac{3.426}{(5.722)^{1.5}} \approx 0.25$$

Conclusion (Example 1)

- Pearson ≈ 0.28 , Moment $g_1 \approx 0.25 \rightarrow$ slight positive skew.
- Bowley = 0 (quartile symmetry) — not uncommon with small samples. Overall: slightly positively skewed (mean–median–mode and moment/Pearson support this).

Exercise 2: grouped data

The number of customer complaints in 40 companies is grouped as follows:

Class (x_j')	Frequency (n_j)
0–10	5
10–20	8
20–30	12
30–40	10
40–50	5

Tasks:

- 1 Calculate the **grouped mean**.
- 2 Estimate the **median and mode**.
- 3 Determine skewness direction using the **mean–median–mode relationship**.
- 4 Calculate the **Pearson's coefficient**, **Bowley's coefficient**, and **moment coefficient**.
- 5 Compare the results and discuss whether the distribution is symmetric, positively skewed, or negatively skewed.

Exercise 2: solution

Answer: Example 2: Grouped Data – Class Table (corrected)

Class interval	x'_j (Midpoint)	n_j	$f_j = n_j/n$	N_j (cum abs)	F_j^* (cum rel)
[0 – 10]	5	5	0.125	5	0.125
(10 – 20]	15	8	0.200	13	0.325
(20 – 30]	25	12	0.300	25	0.625
(30 – 40]	35	10	0.250	35	0.875
(40 – 50]	45	5	0.125	40	1.000

Exercise 2: solution

- **Mean (grouped)**

$$\bar{X} = \sum f_j x'_j = 0.125(5) + 0.200(15) + 0.300(25) + 0.250(35) + 0.125(45) =$$

- **Median** (use median class interpolation with cumulative relative freqs)
median class is the first with $F_j^* \geq 0.5 \rightarrow [20 - 30)$. Apply

$$\text{Med} = L + \frac{0.5 - F_{L-1}^*}{F_L^* - F_{L-1}^*} h$$

where $L = 20$, $F_{L-1}^* = 0.325$, $F_L^* = 0.625$, $h = 10$. Thus

$$\text{Med} = 20 + \frac{0.5 - 0.325}{0.625 - 0.325} \times 10 = 20 + \frac{0.175}{0.300} \times 10 = 20 + 5.8333 = 25.8333$$

Exercise 2: solution

- **Mode** (grouped, using relative frequencies) modal class = [20 – 30) (largest $f_L = 0.30$). Use

$$f^* = f_L - f_{L-1}, \quad f^{**} = f_L - f_{L+1}$$

$$Mo = L + \frac{f^*}{f^* + f^{**}}h$$

Here $f_L = 0.30$, $f_{L-1} = 0.20$, $f_{L+1} = 0.25$ so $f^* = 0.10$, $f^{**} = 0.05$.
Then

$$Mo = 20 + \frac{0.10}{0.10 + 0.05} \times 10 = 20 + \frac{0.10}{0.15} \times 10 = 20 + 6.6667 = 26.6667$$

Exercise 2: solution

$$\bar{X} = 25.50, \quad \text{Med} \approx 25.833, \quad \text{Mo} \approx 26.667$$

So $\bar{X} < \text{Med} < \text{Mo} \rightarrow$ this indicates a slight negative (left) skew.

Central moments (grouped, about grouped mean) Using midpoints and f_j (population central moments):

$$\bullet m_2 = \sum f_j(x'_j - \bar{X})^2 = 144.75$$

$$\Rightarrow s = \sqrt{m_2} = 12.0312$$

$$\bullet m_3 = \sum f_j(x'_j - \bar{X})^3 = -167.25$$

(rounded values shown above; calculations use midpoints and f_j .)

Exercise 2: solution

Answer: Skewness coefficients

- Pearson (using mean & mode):

$$g = \frac{\bar{X} - Mo}{s} = \frac{25.50 - 26.6667}{12.0312} \approx -0.097$$

- Bowley (quartile-based) — compute quartiles by interpolation with F^* :

$$Q_1 \approx 17.5, \quad Q_2 = \text{Med} \approx 25.8333, \quad Q_3 \approx 33.0$$

$$g' = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \approx \frac{(33.0 - 25.8333) - (25.8333 - 17.5)}{(33.0 - 25.8333) + (25.8333 - 17.5)} \approx -0.022$$

- Moment coefficient (Fisher / g_1):

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{-167.25}{(144.75)^{1.5}} \approx -0.096$$

Final interpretation (concise)

- All three numerical measures (Pearson ≈ -0.097 , Bowley ≈ -0.022 , Moment $g_1 \approx -0.096$) are **small in magnitude** and **slightly negative**, and the ordering mean $<$ median $<$ mode also indicates **slight negative skewness**.
- Conclusion: the grouped distribution is **approximately symmetric but with a very small left skew**.

Exercise 3: Graphical Analysis

For both Exercise 1 and Exercise 2:

- 1 Draw a **histogram** (or bar chart for Exercise 2) and a **boxplot**.
- 2 Use the graphs to **visualize skewness**.
- 3 Compare the graphical impression with the calculated coefficients.

Exercise 3.2 (Suggested Answer)

We can prepare the frequency table first:

Table: Frequency Distribution and Midpoints for Monthly Wages

Salary Interval (euros)	f_j	F_j^*	Midpoint x'_j
[600, 900]	0.09	0.09	750
]900, 1000]	0.18	0.27	950
]1000, 1100]	0.35	0.62	1050
]1100, 1200]	0.22	0.84	1150
]1200, 1300]	0.16	1.00	1250

Exercise 3.2 (Suggested Answer)

a) To construct a boxplot, we determine the five-number summary: (Min, Q_1 , Q_2 , Q_3 , Max).

- Min = 600 (lower limit of the first class)
- Max = 1300 (upper limit of the last class)
- grouped quantile formula:

$$q_p = l + \frac{p - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

$$Q_1 = 900 + \frac{0.25 - 0.09}{0.27 - 0.09} \cdot (1000 - 900) = 900 + \frac{0.16}{0.18} \cdot 100 \approx 988.889$$

$$Q_2 = 1000 + \frac{0.50 - 0.27}{0.62 - 0.27} \cdot (1100 - 1000) = 1000 + \frac{0.23}{0.35} \cdot 100 \approx 1065.714$$

$$Q_3 = 1100 + \frac{0.75 - 0.62}{0.84 - 0.62} \cdot (1200 - 1100) = 1100 + \frac{0.13}{0.22} \cdot 100 \approx 1159.091$$

Exercise 3.2 (Suggested Answer)

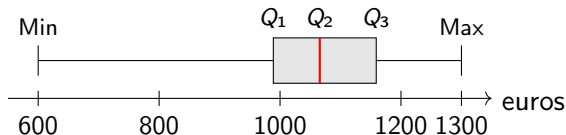


Figure: Boxplot of Monthly Wages

It has a negative skewness because the lower whisker is significantly longer than the upper whisker.

Exercise 3.2 (Suggested Answer)

- b) Calculate the Original Mean ($\bar{x}$): Using the grouped mean formula ($\bar{x} = \sum f_j \cdot x'_j$):

$$\bar{x} = (0.09 \times 750) + (0.18 \times 950) + (0.35 \times 1050) + (0.22 \times 1150) + (0.16 \times 1250)$$

$$\bar{x} = 67.5 + 171.0 + 367.5 + 253.0 + 200.0 = 1059.0$$

$$\bar{x}_{\text{new}} = 1059.0 \times 1.10 = 1164.9 \text{ euros}$$

Exercise 3.2 (Suggested Answer)

- c) To evaluate concentration, we calculate the Gini Index (G) using the discrete formula for grouped data:

$$G = 1 - \frac{\sum_{i=1}^{m-1} Q_i}{\sum_{i=1}^{m-1} P_i}$$

Exercise 3.2 (Suggested Answer)

Table: Initial Data and Class Totals for Monthly Wages

Class (k)	n_j	x'_j	$t_j = n_j x'_j$
[600, 900]	9	750.00	6750.00
]900, 1000]	18	950.00	17100.00
]1000, 1100]	35	1050.00	36750.00
]1100, 1200]	22	1150.00	25300.00
]1200, 1300]	16	1250.00	20000.00
Total	$N = 100$		$T = 105900.00$

Table: Shares, Cumulative Shares, and Gini Index Components

Class	n_j/N	t_j/T	P_i (cum.)	Q_i (cum.)
[600, 900]	0.09000	0.06374	0.09000	0.06374
]900, 1000]	0.18000	0.16157	0.27000	0.22531
]1000, 1100]	0.35000	0.34703	0.62000	0.57234
]1100, 1200]	0.22000	0.23890	0.84000	0.81124
]1200, 1300]	0.16000	0.18876	1.00000	1.00000

Exercise 3.2 (Suggested Answer)

$$\sum_{i=1}^4 Q_i = 0.06374 + 0.22531 + 0.57234 + 0.81124 = 1.67263$$

$$\sum_{i=1}^4 P_i = 0.09 + 0.27 + 0.62 + 0.84 = 1.82$$

$$G = 1 - \frac{1.67263}{1.82} \approx 0.081$$

Exercise 3.4 (Suggested Answer)

a) Salary Distribution Summary Table

Classes (j)	n_j	Midpoint (x'_j)	f_j	F_j^*
0 – 3	500	1.5	0.10	0.10
3 – 6	1500	4.5	0.30	0.40
6 – 9	2500	7.5	0.50	0.90
9 – 12	500	10.5	0.10	1.00
Total	$N = 5000$	—	1.00	—

$$\begin{aligned}\bar{x} &= \frac{\sum_{j=1}^k n_j x'_j}{N} \\ &= \frac{(500 \times 1.5) + (1500 \times 4.5) + (2500 \times 7.5) + (500 \times 10.5)}{5000} \\ &= 6.3.\end{aligned}\quad (2)$$

Exercise 3.4 (Suggested Answer)

Median:

$$Me = l + \frac{0.5 - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

- Lower boundary (l) = 6, Upper boundary (L) = 9
- Class width (h) = $9 - 6 = 3$
- $F^*(l) = F^*(6) = 0.40$
- $F^*(L) = F^*(9) = 0.90$

$$Me = 6 + \frac{0.5 - 0.40}{0.90 - 0.40} \cdot 3 = \mathbf{6.6}$$

Exercise 3.4 (Suggested Answer)

b) The general quantile formula is:

$$Q_p = l + \frac{p - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

- Q_p is the value such that: a proportion p of the observations fall below Q_p , while the remaining proportion $1 - p$ lies above it.
- To find Q_p , we plug in p into the formula.
- The question now asks for the percentage of workers earning more than the mean salary. So the question is: what is the proportion p of the observations **fall below the Mean**, while the remaining proportion $1 - p$ lies above the mean.

To solve p : we can find it using the above formula

$$Mean = l + \frac{p - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

Exercise 3.4 (Suggested Answer)

Since the mean is 6.3, we set $Q_p = 6.3$ and we have

- Lower boundary (I) = 6
- Upper boundary (L) = 9
- Class width (h) = 3
- Cumulative relative frequency at I ($F^*(6)$) = 0.40
- Cumulative relative frequency at L ($F^*(9)$) = 0.90

Substitute these into the formula:

$$6.3 = 6 + \frac{p - 0.40}{0.90 - 0.40} \cdot 3,$$

solve for p , we get $p = 0.45$. So $55\% = 1 - 45\%$ of workers earn more than the mean salary.

Exercise 3.4 (Suggested Answer)

- c) Find the target proportion (p): For the lowest 30%, our target cumulative proportion is $p = 0.30$. We use the Quantile Formula:

$$Q_p = l + \frac{p - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

$$Q_{0.30} = 3 + \frac{0.30 - 0.10}{0.40 - 0.10} \cdot 3 = 3 + \frac{0.20}{0.30} \cdot 3 = 5$$

$$N = 5000 \times 30\% = 1500.$$

Exercise 3.4 (Suggested Answer)

So we can construct the following table:

Table: Salary Distribution Summary Table (Lowest 30%)

Classes (j)	n_j	x'_j	f_j	F_j^*
0 – 3	500	1.5	0.333	0.333
3 – 5	1000	4.0	0.667	1.000
Total	$N = 1500$	—	1.000	—

$$\bar{x}_1 = \frac{(500 \times 1.5) + (1000 \times \frac{3+5}{2})}{1500} = \frac{750 + 4000}{1500} = \frac{4750}{1500} \approx 3.167$$

Exercise 3.4 (Suggested Answer)

- d) The highest 30% starts at the 70th percentile, so our lower boundary proportion is $p = 0.70$

$$Q_{0.70} = l + \frac{0.70 - F^*(l)}{F^*(L) - F^*(l)} \cdot h$$

- Target class is $[6, 9)$ because $p = 0.70$ lies between $F^*(6) = 0.40$ and $F^*(9) = 0.90$.
- $l = 6$, $L = 9$, $h = 3$, $F^*(l) = 0.40$, $F^*(L) = 0.90$.

$$Q_{0.70} = 6 + \frac{0.70 - 0.40}{0.90 - 0.40} \cdot 3 = 6 + \frac{0.30}{0.50} \cdot 3 = 7.8$$

Exercise 3.4 (Suggested Answer)

So we can construct the following table:

Table: Salary Distribution Summary Table (Highest 30%)

Classes (j)	n_j	x'_j	f_j	F_j^*
7.8 – 9	1000	8.4	0.667	0.667
9 – 12	500	10.5	0.333	1.000
Total	$N = 1500$	—	1.000	—

$$\bar{x}_2 = \frac{(1000 \times 8.4) + (500 \times 10.5)}{1500} = \frac{8400 + 5250}{1500} = \frac{13650}{1500} = 9.1$$

Exercise 3.5 (Suggested Answer)

a) $s_{ST}^2 = 4.41$, $s_{ECO}^2 = 3.69$

b) $CV_{ST} = 16.535$, $CV_{ECO} = 13.820$

Exercise 3.6 (Suggested Answer)

a)

$$\bar{x}_A = \frac{7.8 + 7.0 + 8.2 + 7.6 + 6.9 + 7.7 + 7.2 + 7.8 + 7.3 + 7.5}{10} = \frac{75}{10} = \mathbf{7.5 \text{ litres}}$$

$$\bar{x}_B = \frac{7.0 + 7.1 + 7.1 + 7.3 + 7.2 + 7.0 + 7.8 + 8.0 + 7.0 + 7.0}{10} = \frac{72.5}{10} = \mathbf{7.25 \text{ litres}}$$

Exercise 3.6 (Suggested Answer)

- b)** Note: The Normal Stem (Lower Half): Holds leaves from 0 to 4. The Star Stem ($\star$) (Upper Half): Holds leaves from 5 to 9.

Leaves – A		Leaves – B
9	6 $\star$	
0 2 3	7	0 0 0 0 1 1 2 3
5 6 7 8 8	7 $\star$	8
2	8	0

- Brand A (Left): Values are relatively spread across the center classes, showing a relatively balanced distribution around the middle values.
- Heavily concentrated at the lower end (four 7.0 values and multiple 7.1 values), indicating a strong positive skewness .

- c)** $F^*(6.9) = 0, F^*(7.3) = 0.4, F^*(7.8) = 0.9, F^*(8.2) = 1; 0.6.$

THANKS!

Questions?