Statistics and Probability
Detailed Correction – Exercise Set 1
Exercise 1
Exercise 2
Exercise 3

One-Dimensional Statistical Series — Detailed Correction

Statistics and Probability — Detailed Correction – Exercise Set 1

Focus on the Way of Thinking, Explicit Substitutions, and Step-by-Step Course Execution

M.Guellati · Semester 1 · 2026/2027

How to Use This Correction

Each solution follows a strict four-stage analytical structure designed to train exam intuition: **Method** (identifies the exact course technique) ⟶ **Why we do this** (explains the mental train of thought and algebraic clues) ⟶ **Execution** (complete line-by-line derivation with zero skipped steps and explicit **Sub** callouts) ⟶ **Pitfall** / **Sanity Check** (where students lose points and how to verify your answer).

Difficulty: ★☆☆☆☆ Direct ★★☆☆☆ Standard ★★★☆☆ Exam level ★★★★☆ Tricky ★★★★★ Synthesis

Exercise 1

The triglyceride level (in g/L) was measured in 250 men aged 20–30 years:

Triglycerides (g/L)	[0.4, 0.6[	[0.6, 0.8[	[0.8, 1.0[	[1.0, 1.2[	[1.2, 1.4[	[1.4, 1.6[
Number of men	5	32	86	89	32	6

1. Calculate the relative frequencies and the cumulative ascending frequencies. ★☆☆☆☆

Method *Relative and cumulative frequencies of a grouped continuous variable*

> ***Why we do this (Way of Thinking):***
> A table of classes gives only counts n_i. A count alone says nothing about the *share* of the population it represents, so the first move is always to find the total N and divide: $f_i = n_i/N$. For the cumulative columns we ask a different question: "how many men are *below* this upper bound?" (ascending) or "how many are *above* this lower bound?" (descending). Both are simply running totals, one from the top of the table, one from the bottom. We compute the descending ones now as well, because Question 4 needs them.

Execution: First the total:

$N = 5+32+86+89+32+6$: $\quad 5+32=37,\ 37+86=123,\ 123+89=212,\ 212+32=244,\ 244+6=250.$

Relative frequencies $f_i = \dfrac{n_i}{250}$:

$$\frac{5}{250}=0.020,\quad \frac{32}{250}=0.128,\quad \frac{86}{250}=0.344,\quad \frac{89}{250}=0.356,\quad \frac{32}{250}=0.128,\quad \frac{6}{250}=0.024.$$

Ascending counts (running total from the first class): $5,\ 5+32=37,\ 37+86=123,\ 123+89=212,\ 212+32=244,\ 244+6=250$, and then $F_i^{\nearrow} = n_i^{\nearrow}/250$:

$$\frac{5}{250}=0.020,\ \frac{37}{250}=0.148,\ \frac{123}{250}=0.492,\ \frac{212}{250}=0.848,\ \frac{244}{250}=0.976,\ \frac{250}{250}=1.$$

Descending counts (running total from the last class): $6,\ 6+32=38,\ 38+89=127,\ 127+86=213,\ 213+32=245,\ 245+5=250$, and $F_i^{\searrow} = n_i^{\searrow}/250$:

$$\frac{250}{250}=1,\ \frac{245}{250}=0.980,\ \frac{213}{250}=0.852,\ \frac{127}{250}=0.508,\ \frac{38}{250}=0.152,\ \frac{6}{250}=0.024.$$

Class (g/L)	n_i	f_i	$n_i^{\nearrow}$	$F_i^{\nearrow}$	$n_i^{\searrow}$	$F_i^{\searrow}$
$[0.4, 0.6[$	5	0.020	5	0.020	250	1.000
$[0.6, 0.8[$	32	0.128	37	0.148	245	0.980
$[0.8, 1.0[$	86	0.344	123	0.492	213	0.852
$[1.0, 1.2[$	89	0.356	212	0.848	127	0.508
$[1.2, 1.4[$	32	0.128	244	0.976	38	0.152
$[1.4, 1.6[$	6	0.024	250	1.000	6	0.024
Total	250	1				

$$N = 250, \qquad \sum_i f_i = 1$$

► **Common Pitfall:**
Dividing by the number of classes (6) or by the largest count instead of by $N = 250$. Another classic slip: attaching the ascending cumulative value to the *lower* bound of its class (it belongs to the *upper* bound), and the descending one to the upper bound (it belongs to the lower bound).

✓ **Sanity Check:**
$\sum f_i = 0.020 + 0.128 + 0.344 + 0.356 + 0.128 + 0.024 = 1$; the last ascending value must be 1 and the first descending value must be 1. Cross-check: each descending value equals 1 − the previous ascending value (for instance $0.508 = 1 - 0.492$).

2. Draw the histogram and its frequency polygon. ★★☆☆☆

Method ***Histogram with equal class widths and frequency polygon***

Why we do this (Way of Thinking):
In a histogram the *area* of each rectangle must be proportional to the frequency of its class. Here every class has the same width $a = 0.2$, so we can use the counts n_i directly as heights: equal widths mean area is proportional to height, nothing to correct. The rectangles touch each other, because the variable is continuous and the classes follow one another without gaps. For the polygon we want a broken line that follows the shape of the distribution: we join the *midpoints* of the tops of the rectangles, and we close it on the axis at the centres of the two empty classes before the first and after the last (here 0.3 and 1.7), so that the polygon keeps the same area as the histogram.

Execution: Class centres $c_i = \dfrac{\text{lower} + \text{upper}}{2}$: $\dfrac{0.4+0.6}{2} = 0.5,\ 0.7,\ 0.9,\ 1.1,\ 1.3,\ 1.5$. Rectangles: base $[a, a+0.2[$, height n_i. Polygon vertices:

$$(0.3, 0),\ (0.5, 5),\ (0.7, 32),\ (0.9, 86),\ (1.1, 89),\ (1.3, 32),\ (1.5, 6),\ (1.7, 0).$$

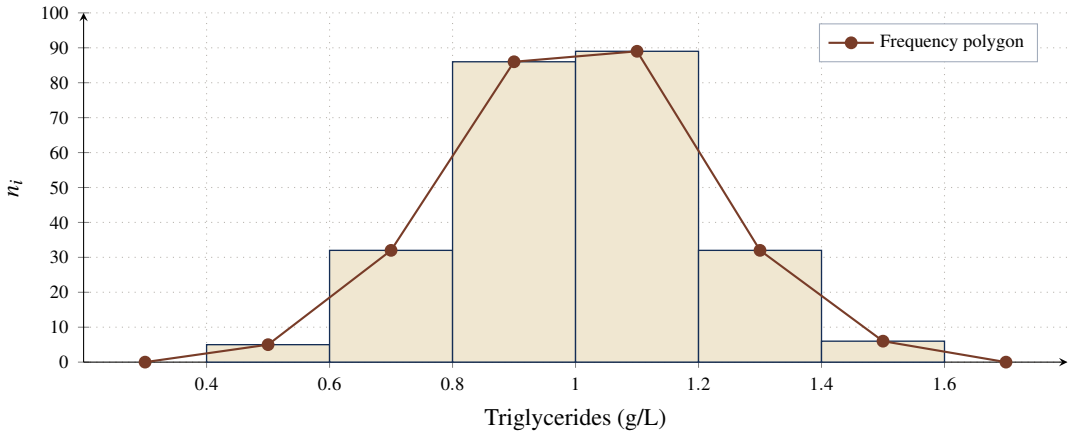


Figure 1: Histogram and frequency polygon of the triglyceride level.

> **► Common Pitfall:**
> Leaving gaps between the rectangles (that is a bar chart, not a histogram), joining the *corners* of the rectangles instead of the midpoints of their tops, or forgetting to bring the polygon back to the axis at both ends.

> **✓ Sanity Check:**
> The heights must read $5, 32, 86, 89, 32, 6$ and the tallest rectangle must be over $[1.0, 1.2[$. The polygon must pass through the middle of the top of every rectangle and touch the axis at 0.3 and 1.7.

3. Determine the modal class graphically and by calculation.

★★★☆☆

Method ***Modal class, graphical construction and interpolation formula for the mode***

> ***Why we do this (Way of Thinking):***
> Since all the classes have the same width, the modal class is simply the one with the largest count: $[1.0, 1.2[$ (89 men). But the mode is a *value*, and the true peak of the distribution sits somewhere inside that class, pulled towards the neighbour that is the larger one. Look at the neighbours: on the left 86 is almost as large as 89, on the right 32 is much smaller. So the peak must lean towards the *left* end of the class, and we expect a mode very close to 1.0. The graphical construction (two crossing segments) and the formula both encode exactly this "pull" of the neighbours.

Execution: Graphically. In the histogram, join the upper-left corner of the modal rectangle, $(1.0, 89)$, to the upper-left corner of the following rectangle, $(1.2, 32)$; then join the upper-right corner of the modal rectangle, $(1.2, 89)$, to the upper-right corner of the preceding rectangle, $(1.0, 86)$. The abscissa of the intersection point is the mode.

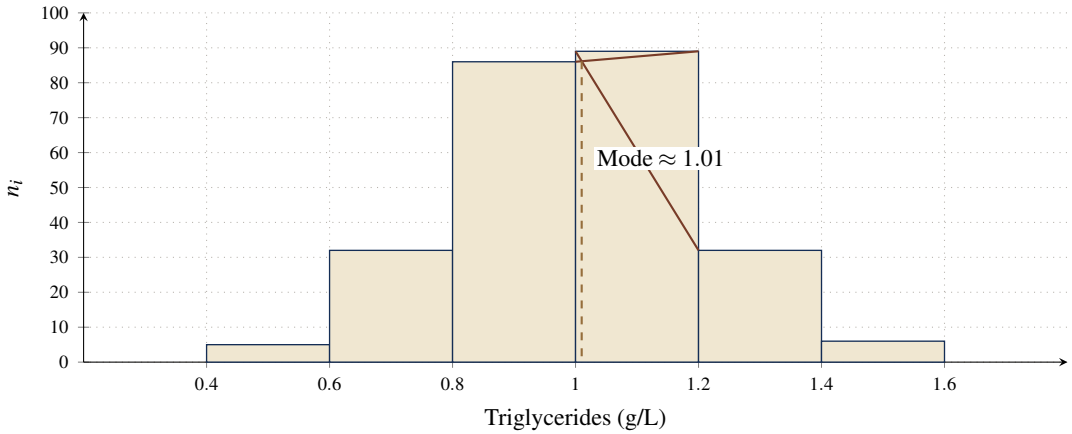


Figure 2: Graphical determination of the mode.

By calculation. Modal class $[1.0, 1.2[$, lower bound 1.0, width 0.2:

$$\text{Mode} = 1.0 + 0.2 \times \frac{89-86}{(89-86)+(89-32)} = 1.0 + 0.2 \times \frac{3}{3+57}$$
$$= 1.0 + 0.2 \times \frac{3}{60} = 1.0 + 0.2 \times 0.05 = 1.0 + 0.01.$$

$$\boxed{\text{Mode} = 1.01 \text{ g/L}}$$

Most of men have 1.01 g/L in their blood.

> **► Common Pitfall:**
> Swapping the two neighbours in the formula (using $89-32$ on top and $89-86$ at the bottom gives a mode near the *right* end, which contradicts the picture), or forgetting to multiply by the class width 0.2.

> **✓ Sanity Check:**
> The mode must lie inside the modal class: $1.0 \leq 1.01 < 1.2$ ✓. It must also agree with the graphical construction (the segments cross at $x \approx 1.01$, very close to the left end, as predicted).

4. Plot the cumulative frequency curves (ascending and descending) on the same graph. Determine the median graphically, calculate it, and interpret the value. ★★★☆☆

Method ***Cumulative frequency curves and the median by linear interpolation***

> ***Why we do this (Way of Thinking):***
> The median is the value that splits the population in two halves: 50 % of the men are below it and 50 % above it. In terms of the cumulative curve this is the abscissa where the ascending frequency reaches 0.5. The descending curve gives "the share above x", so at the median it *also* equals 0.5: that is why the two curves cross exactly at the height 0.5, and the abscissa of the crossing is the median. Looking at the table, $F^{\nearrow}$ is 0.492 at 1.0 and 0.848 at 1.2: the value 0.5 is just above 0.492, so the median is inside $[1.0, 1.2[$ and very close to its left end.

Execution: Graph. The ascending curve passes through the points (upper bound, $F^{\nearrow}$): $(0.4, 0), (0.6, 0.02), (0.8, 0.148), (1.0, 0.492), (1.2, 0.848), (1.4, 0.976), (1.6, 1)$. The descending curve passes through (lower bound, $F^{\searrow}$): $(0.4, 1), (0.6, 0.98), (0.8, 0.852), (1.0, 0.508), (1.2, 0.152), (1.4, 0.024), (1.6, 0)$.

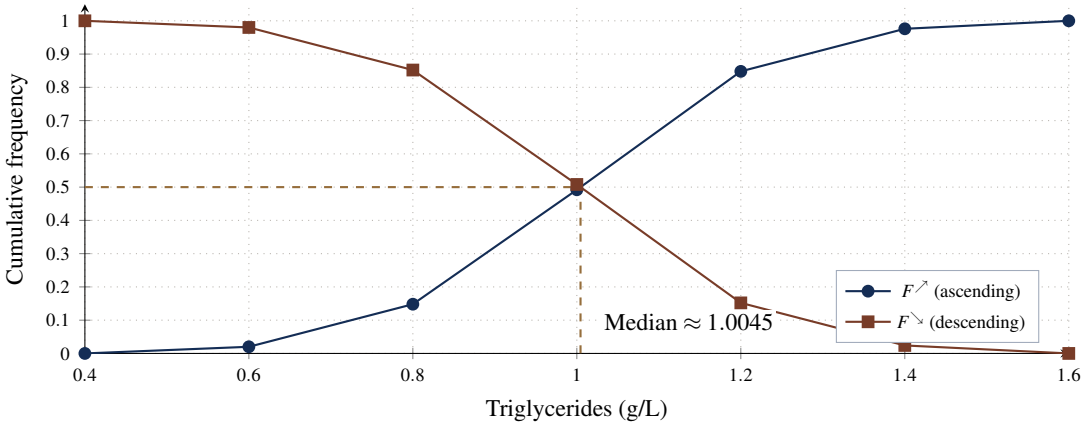


Figure 3: Ascending and descending cumulative curves; the median is read at the crossing, at height 0.5.

Calculation. Half of the population: $\dfrac{N}{2} = \dfrac{250}{2} = 125$. Looking at the ascending counts: $123 < 125 \leq 212$, so the median class is $[1.0, 1.2[$ (lower bound 1.0, width 0.2, frequency $f = 0.356$, ascending frequency of the *previous* class 0.492):

$$\text{Median} = 1.0 + 0.2 \times \frac{0.5 - 0.492}{0.356} = 1.0 + 0.2 \times \frac{0.008}{0.356} = 1.0 + 0.2 \times 0.0225 \approx 1.0 + 0.0045.$$

$$\boxed{\text{Median} \approx 1.0045 \text{ g/L}}$$

Interpretation. 50 % of the men have a triglyceride level of at most 1.0045 g/L (about 1.00 g/L), and the other 50 % have a level of at least that value.

> **► Common Pitfall:**
> Plotting the ascending points at the class centres or at the lower bounds (they belong to the *upper* bounds), and, in the formula, using the cumulative frequency of the median class (0.848) instead of that of the class *just before* it (0.492).

> **✓ Sanity Check:**
> Between $x = 1.0$ and $x = 1.2$ the ascending curve rises by 0.356 over 0.2, i.e. with slope 1.78. At $x = 1.0045$ it gives $0.492 + 1.78 \times 0.0045 \approx 0.5$ ✓. The calculated value (1.0045) must match the crossing read on the graph.

Exercise 2

The mass (in grams) of eggs from a farm was recorded:

Egg mass (g)	$[20,30[$	$[30,40[$	$[40,50[$	$[50,60[$	$[60,70[$	$[70,80[$	$[80,90[$
Number of eggs	3	51	74	112	92	62	6

1. Define the population, the variable, its nature, and its modalities. ★☆☆☆☆

Method *Vocabulary of descriptive statistics*

> ***Why we do this (Way of Thinking):***
> Four questions, four answers. *Who or what is observed?* That gives the population (and its unit). *What is measured on each of them?* That is the variable. *Is it a label or a number, and if it is a number, can it take any value in a range?* That is its nature. *Which values (or groups of values) does it take?* Those are the modalities. Here the table counts eggs and sorts them by mass, so the unit is an egg and the variable is its mass.

Execution:

- **Population:** the eggs of the farm that were recorded; the statistical unit is one egg. Their number is $N = 3+51+74+112+92+62+6 = 400$.
- **Variable:** the mass of an egg, in grams.
- **Nature:** the mass can take any real value in a range (for instance 52.37 g), and the data are grouped into classes: it is a *continuous quantitative* variable.
- **Modalities:** the classes $[20,30[$, $[30,40[$, $[40,50[$, $[50,60[$, $[60,70[$, $[70,80[$, $[80,90[$.

variable: egg mass (g) — continuous quantitative

> **► Common Pitfall:**
> Giving "the number of eggs" as the variable (it is the *frequency*, not the variable), or calling the mass *discrete* because it is presented in classes.

> **✓ Sanity Check:**
> Test the nature: can we meaningfully compute an average of the variable, and can an intermediate value such as 52.37 g occur? Yes to both, so it is quantitative and continuous.

2. Complete the table with class centers, cumulative frequencies (ascending and descending), and relative cumulative frequencies (ascending and descending). ★★☆☆☆

Method *Complete frequency table of a grouped variable*

> ***Why we do this (Way of Thinking):***
> Every new column is built from the previous one, so we go in a fixed order: centre of each class, relative frequency $f_i = n_i/N$, then running totals. Ascending totals count the eggs lighter than the upper bound of the class, descending totals count the eggs heavier than its lower bound. The relative cumulative frequencies are just those totals divided by $N = 400$.

Execution: Centres $c_i = \dfrac{\text{lower}+\text{upper}}{2}$: $\dfrac{20+30}{2} = 25,\ 35,\ 45,\ 55,\ 65,\ 75,\ 85$.
Relative frequencies $f_i = \dfrac{n_i}{400}$: $\frac{3}{400} = 0.0075$, $\frac{51}{400} = 0.1275$, $\frac{74}{400} = 0.185$, $\frac{112}{400} = 0.28$, $\frac{92}{400} = 0.23$, $\frac{62}{400} = 0.155$, $\frac{6}{400} = 0.015$.
Ascending counts: $3,\ 3+51 = 54,\ 54+74 = 128,\ 128+112 = 240,\ 240+92 = 332,\ 332+62 = 394,\ 394+6 = 400$.
Descending counts: $400,\ 400-3 = 397,\ 397-51 = 346,\ 346-74 = 272,\ 272-112 = 160,\ 160-92 = 68,\ 68-62 = 6$.
Relative ascending $F_i^{\nearrow} = \dfrac{n_i^{\nearrow}}{400}$: $0.0075,\ 0.135,\ 0.32,\ 0.6,\ 0.83,\ 0.985,\ 1$. Relative descending $F_i^{\searrow} = \dfrac{n_i^{\searrow}}{400}$: $1,\ 0.9925,\ 0.865,\ 0.68,\ 0.4,\ 0.17,\ 0.015$.

Class (g)	c_i	n_i	$n_i^{\nearrow}$	$n_i^{\searrow}$	f_i	$F_i^{\nearrow}$	$F_i^{\searrow}$
$[20,30[$	25	3	3	400	0.0075	0.0075	1.0000
$[30,40[$	35	51	54	397	0.1275	0.1350	0.9925
$[40,50[$	45	74	128	346	0.1850	0.3200	0.8650
$[50,60[$	55	112	240	272	0.2800	0.6000	0.6800
$[60,70[$	65	92	332	160	0.2300	0.8300	0.4000
$[70,80[$	75	62	394	68	0.1550	0.9850	0.1700
$[80,90[$	85	6	400	6	0.0150	1.0000	0.0150
Total		400			1		

$$N = 400, \qquad \sum_i f_i = 1$$

► Common Pitfall:
Mixing up the two directions: the ascending column starts with the *smallest* class and ends at N, the descending column starts at N and ends with the count of the *largest* class. Also, computing a centre from the width instead of from the two bounds.

✓ Sanity Check:
$\sum f_i = 0.0075+0.1275+0.185+0.28+0.23+0.155+0.015 = 1$; $n^{\nearrow}$ ends at 400; $n^{\searrow}$ starts at 400 and ends at $6 = n_7$; and for every row $F_i^{\searrow} = 1 - F_{i-1}^{\nearrow}$ (for instance $0.865 = 1 - 0.135$).

3. Draw the histogram of frequencies.

★★☆☆☆

Method *Histogram with equal class widths*

Why we do this (Way of Thinking):
All seven classes have the same width ($a = 10$ g), so the area of a rectangle is proportional to its height, and we can plot the counts n_i directly. No correction is needed here (it will be needed in Question 6, where two classes become twice as wide).

Execution: Draw one rectangle per class, on the base $[a, a+10[$, with heights 3, 51, 74, 112, 92, 62, 6.

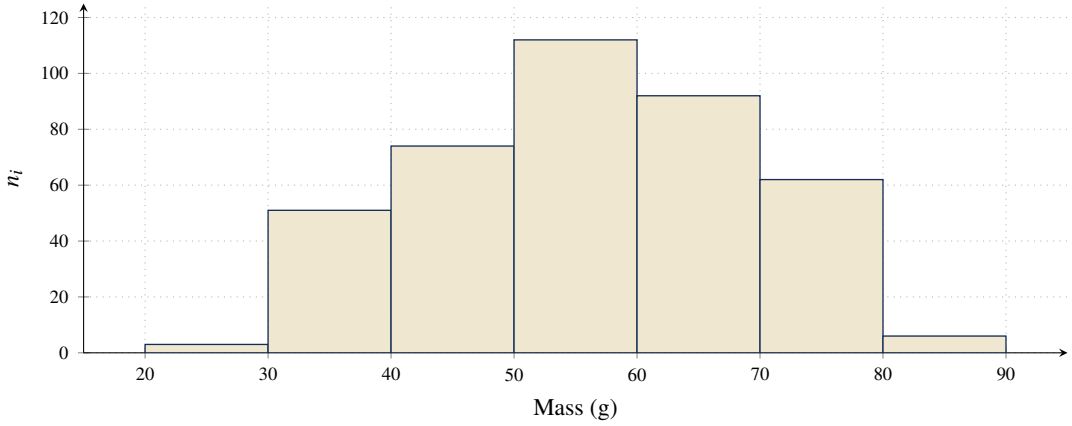


Figure 4: Histogram of the egg masses.

> **► Common Pitfall:**
> Drawing separated bars, or choosing a vertical scale that is not uniform. Also plotting the centres c_i on the horizontal axis as points: the horizontal axis must show the *class bounds*.

> **✓ Sanity Check:**
> The tallest rectangle (112) is over $[50,60[$ and the shape is roughly symmetric around it, consistent with the counts 74 and 92 on either side.

4. Estimate the number of eggs weighing between 45 g and 63 g. Indicate this on the histogram. ★★★☆☆

Method ***Estimation by linear interpolation inside classes (two equivalent methods)***

> ***Why we do this (Way of Thinking):***
> The values 45 g and 63 g cut two classes in the middle, and the table does not tell us where exactly the eggs of a class lie. The only reasonable assumption is that the eggs are spread *evenly* inside each class: then a part of a class that covers a fraction $\frac{x}{a}$ of its width contains the same fraction of its eggs (on the histogram, this is just the part of the area of the rectangle that is covered). From 45 to 63 we meet: the right half of $[40,50[$ (from 45 to 50), the whole class $[50,60[$, and the left 3 tenths of $[60,70[$ (from 60 to 63). The second method rewrites the same idea with the cumulative function F: the share of eggs between two masses is $F(63)-F(45)$.

Execution: Method 1 (proportion of each class).

$$\begin{aligned}
\text{in } [45,50[:\ & 74\times\frac{50-45}{10}=74\times 0.5=37,\\
\text{in } [50,60[:\ & 112 \text{ (whole class)},\\
\text{in } [60,63[:\ & 92\times\frac{63-60}{10}=92\times 0.3=27.6.
\end{aligned}$$

$$n=37+112+27.6=176.6\approx 177 \text{ eggs} \qquad (176.6/400\approx 44\,\% \text{ of the eggs}).$$

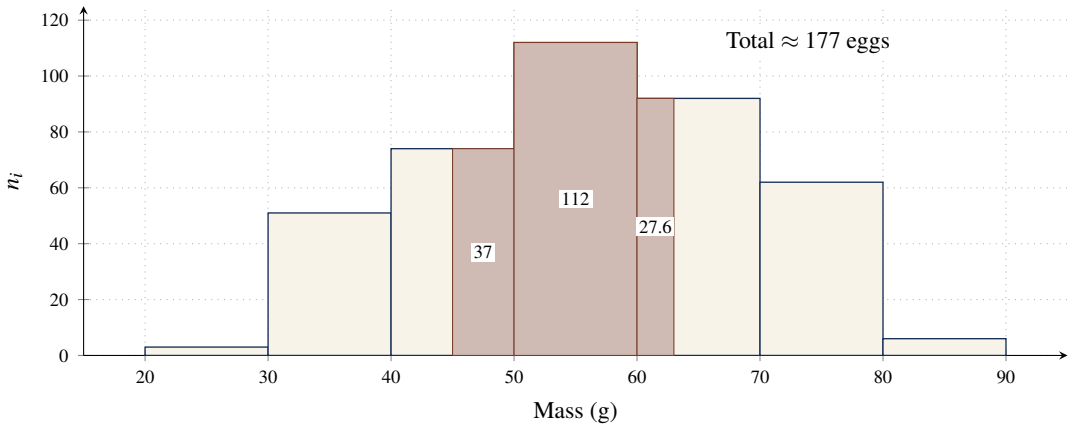


Figure 5: The shaded area corresponds to the eggs between 45 g and 63 g.

Method 2 (using the CDF). The cumulative distribution function $F(x)$ is the proportion of eggs with a mass $\leq x$. It is linear between two class bounds a and b (ogive):

$$F(x)=F(a)+\frac{x-a}{b-a}\big(F(b)-F(a)\big), \qquad a\leq x\leq b,$$

and the number of eggs between x_1 and x_2 is $n = N\big(F(x_2) - F(x_1)\big)$.

$$F(45) = F(40) + \frac{45-40}{50-40}\big(F(50) - F(40)\big) = 0.135 + 0.5\,(0.32 - 0.135) = 0.135 + 0.0925 = 0.2275,$$
$$F(63) = F(60) + \frac{63-60}{70-60}\big(F(70) - F(60)\big) = 0.6 + 0.3\,(0.83 - 0.6) = 0.6 + 0.069 = 0.669.$$

$$n = 400 \times \big(F(63) - F(45)\big) = 400 \times (0.669 - 0.2275) = 400 \times 0.4415 = 176.6$$

$$n \approx 177 \text{ eggs}$$

► **Common Pitfall:**
Counting whole classes (here $74 + 112 + 92 = 278$ eggs, a big overestimate), or using the wrong width for the partial class: from 45 to 50 is 5 g wide (half of the class), not $45 - 40$.

✓ **Sanity Check:**
The estimate must lie between the eggs of the class entirely inside the interval (112) and the eggs of all classes touched by it (278): $112 < 177 < 278$ ✓. Both methods give the same number, 176.6.

5. Determine the modal class and the median class. Calculate the mean mass. Interpret these values. ★★★☆☆

Method *Mode and median by interpolation; mean from the class centres*

Why we do this (Way of Thinking):
Three different ways to say "a typical egg". The *mode* is where the histogram peaks, the *median* is the mass that splits the eggs in two halves, and the *mean* is the balance point. For the first two we first find the class (largest count for the mode, first class whose ascending frequency passes 0.5 for the median), then refine inside the class with the same interpolation idea as before. For the mean, we do not know individual masses, so each egg of a class is represented by the centre of its class.

Execution: Modal class. The widths are equal, so it is the class with the largest count: $[50, 60[$ (112 eggs).

$$\text{Mode} = 50 + 10 \times \frac{112 - 74}{(112 - 74) + (112 - 92)} = 50 + 10 \times \frac{38}{38 + 20} = 50 + 10 \times \frac{38}{58} = 50 + 6.5517 \approx 56.55 \text{ g}$$

Median class. $\dfrac{N}{2} = 200$ and $128 < 200 \leq 240$, so the median class is $[50, 60[$ (previous ascending frequency 0.32, class frequency 0.28):

$$\text{Median} = 50 + 10 \times \frac{0.5 - 0.32}{0.28} = 50 + 10 \times \frac{0.18}{0.28} = 50 + 10 \times 0.6429 \approx 56.43 \text{ g}$$

Mean. Products $n_i c_i$: $3 \times 25 = 75$, $51 \times 35 = 1785$, $74 \times 45 = 3330$, $112 \times 55 = 6160$, $92 \times 65 = 5980$, $62 \times 75 = 4650$, $6 \times 85 = 510$. Their sum: $75 + 1785 = 1860$, $1860 + 3330 = 5190$, $5190 + 6160 = 11350$, $11350 + 5980 = 17330$, $17330 + 4650 = 21980$, $21980 + 510 = 22490$. Then

$$\bar{x} = \frac{1}{N}\sum_i n_i c_i = \frac{22490}{400} = 56.225 \text{ g}$$

$$\text{Mode} \approx 56.55 \text{ g}, \quad \text{Median} \approx 56.43 \text{ g}, \quad \bar{x} = 56.225 \text{ g}$$

Interpretation. The most frequent eggs weigh about 56.5 g; half of the eggs weigh less than 56.43 g and half more; the average egg weighs 56.2 g. Since Mode $\approx$ Median $\approx \bar{x} \approx 56$ g, the distribution is (approximately) symmetric.

> **► Common Pitfall:**
> Taking the centre of the modal class (55) as the mode without using the neighbours, using the ascending frequency of the median class (0.6) instead of that of the class *before* it (0.32) in the median formula, or computing the mean from the class bounds instead of the class centres.

> **✓ Sanity Check:**
> The mode and the median must both fall inside their class $[50,60[$ ✓, and the mean must lie between the smallest and largest centres (25 and 85) ✓. Three nearly equal values are what a symmetric histogram looks like.

6. Regroup the classes into the following intervals: $[20,40[$, $[40,50[$, $[50,60[$, $[60,70[$, $[70,90[$. Draw the new frequency histogram. ★★★★☆

Method ***Histogram with unequal class widths (corrected frequency)***

> ***Why we do this (Way of Thinking):***
> When we merge $[20,30[$ with $[30,40[$ and $[70,80[$ with $[80,90[$, two classes become *twice* as wide as the others. If we simply plotted their frequencies as heights, their rectangles would have twice the area they deserve and the picture would exaggerate the weight of these classes. The rule to remember is: *area must stay proportional to frequency*. So the height of a wide rectangle must be divided by its relative width: we correct each frequency with $f_i^{cor} = f_i \times \dfrac{a_{\min}}{a_i}$, where $a_{\min} = 10$ is the narrowest width.

Execution: Merged counts: $[20,40[$: $3+51=54$; $[70,90[$: $62+6=68$; the other classes are unchanged. Merged frequencies: $f = \frac{54}{400} = 0.135$ and $f = \frac{68}{400} = 0.17$. Corrections: $[20,40[$ has $a=20$, so $f^{cor} = 0.135 \times \frac{10}{20} = 0.0675$; $[70,90[$ has $a=20$, so $f^{cor} = 0.17 \times \frac{10}{20} = 0.085$; for the three classes of width 10, $f^{cor} = f$.

New class	a_i	n_i	f_i	f_i^{cor}	n_i^{cor}
$[20,40[$	20	54	0.135	0.0675	27
$[40,50[$	10	74	0.185	0.185	74
$[50,60[$	10	112	0.280	0.280	112
$[60,70[$	10	92	0.230	0.230	92
$[70,90[$	20	68	0.170	0.085	34
Total		400	1		

Draw the rectangles with heights f_i^{cor}. The polygon of f_i^{cor} passes through the points

$$(15,0),\ (25,0.0675),\ (30,0.0675),\ (35,0.0675),\ (45,0.185),\ (55,0.28),$$
$$(65,0.23),\ (75,0.085),\ (80,0.085),\ (85,0.085),\ (95,0).$$

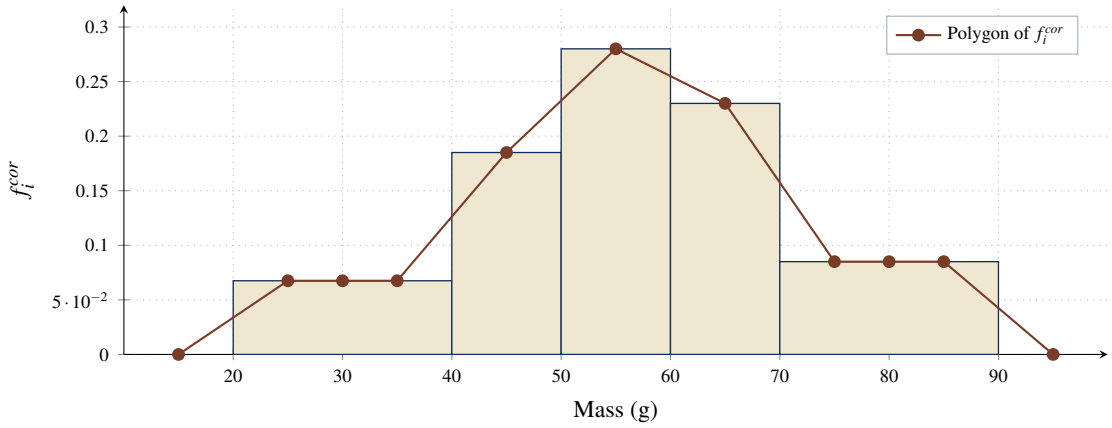


Figure 6: New histogram (corrected frequencies) and its polygon.

> **► Common Pitfall:**
> Plotting f_i (or n_i) as heights for the wide classes without correcting them, or, worse, *multiplying* instead of dividing: it is the *wide* classes whose heights are halved.

> **✓ Sanity Check:**
> Area check: each rectangle represents $f_i^{cor} \times \dfrac{a_i}{a_{\min}}$ units of frequency, and
>
> $$0.0675 \times 2 + 0.185 + 0.28 + 0.23 + 0.085 \times 2 = 0.135 + 0.185 + 0.28 + 0.23 + 0.17 = 1 \checkmark$$

Exercise 3

The distribution of daily wages (in 10^2 DA) for a group of employees is given below. However, one frequency is missing.

Salary class (10^2 DA)	$[5,6[$	$[6,7[$	$[7,8[$	$[8,9[$	$[9,10[$
Number of employees	60	80	?	120	50

1. What is the variable and its nature?

★☆☆☆☆

Method *Vocabulary of descriptive statistics*

> ***Why we do this (Way of Thinking):***
> We look at what is measured on each employee and at the shape of the table: a money amount per day, given in classes of equal width, so a number that can take any value in a range.

Execution: The statistical unit is an employee and the variable is his **daily wage** (in 10^2 DA). It is measured as a number and can take any value inside a range, so it is *quantitative continuous*.

daily wage (in 10^2 DA) — continuous quantitative

> **► Common Pitfall:**
> Saying "the number of employees" (that is the frequency), or calling the wage discrete because it is expressed in whole DA.

> **✓ Sanity Check:**
> Intermediate values such as 7.35 (i.e. 735 DA) are meaningful and the data are grouped into classes: quantitative and continuous ✓.

2. Knowing that the modal class is $[7,8[$ and the calculated mode is 761.1 DA, determine the missing frequency. ★★★★☆

Method *Mode formula used backwards (unknown frequency)*

> ***Why we do this (Way of Thinking):***
> Usually we are given the counts and we compute the mode. Here the roles are reversed: the mode is given and one count, m, is unknown. The mode formula contains m (the modal class count) and the two neighbours 80 and 120, so it becomes an *equation in* m. The trap is units: the table is in 10^2 DA, so the mode 761.1 DA must be written 7.611 before it can be compared with the class $[7,8[$. After that it is only a linear equation: multiply by the denominator, expand, collect the terms in m.

Execution:

Sub **Let m be the missing frequency of $[7,8[$. Units: 761.1 DA $= 7.611$ (in 10^2 DA).**

The modal class is $[7,8[$ (lower bound 7, width 1):

$$\text{Mode} = 7 + 1 \times \frac{m-80}{(m-80)+(m-120)} = 7.611 \implies \frac{m-80}{2m-200} = 7.611 - 7 = 0.611.$$

Multiply both sides by $2m-200$ and expand:

$$m - 80 = 0.611\,(2m-200) = 1.222\,m - 122.2.$$

Collect the terms in m on one side and the constants on the other:

$$m - 1.222\,m = 80 - 122.2 \implies -0.222\,m = -42.2 \implies m = \frac{42.2}{0.222} \approx 190.09.$$

A count is a whole number, so

$$\boxed{m = 190}$$

The total is $N = 60+80+190+120+50 = 500$ employees.

Class (10^2 DA)	c_i	n_i	f_i	$n_i^{\nearrow}$	$F_i^{\nearrow}$	$n_i c_i$	$n_i c_i^2$
$[5,6[$	5.5	60	0.12	60	0.12	330	1815
$[6,7[$	6.5	80	0.16	140	0.28	520	3380
$[7,8[$	7.5	190	0.38	330	0.66	1425	10687.5
$[8,9[$	8.5	120	0.24	450	0.90	1020	8670
$[9,10[$	9.5	50	0.10	500	1.00	475	4512.5
Total		500	1			3770	29065

> **► Common Pitfall:**
> Forgetting the unit conversion (writing 761.1 instead of 7.611 makes the equation meaningless), swapping the two neighbours (80 and 120) in the formula, or leaving $m = 190.09$ as a non-integer count.

> **✓ Sanity Check:**
> Plug $m = 190$ back: $\dfrac{190-80}{(190-80)+(190-120)} = \dfrac{110}{180} = 0.6111$, so Mode $= 7.6111$, i.e. 761.1 DA ✓.
> Also, 190 is indeed larger than its neighbours 80 and 120, so $[7,8[$ really is the modal class.

3. Draw the frequency histogram. Verify the mode graphically.

★★★☆☆

Method *Histogram with equal class widths and graphical construction of the mode*

> ***Why we do this (Way of Thinking):***
> All classes have width 1, so the counts n_i are the heights. For the mode, we repeat the construction of Exercise 1: the segments joining the corners of the modal rectangle to the corners of its two neighbours cross above the mode. Since the right neighbour (120) is larger than the left one (80), the crossing point must lie in the *right* half of $[7,8[$.

Execution: Heights 60, 80, 190, 120, 50. Segment 1: upper-left corner of the modal rectangle $(7,190)$ to the upper-left corner of the next rectangle $(8,120)$. Segment 2: upper-right corner of the modal rectangle $(8,190)$ to the upper-right corner of the previous rectangle $(7,80)$. They cross at $x \approx 7.61$.

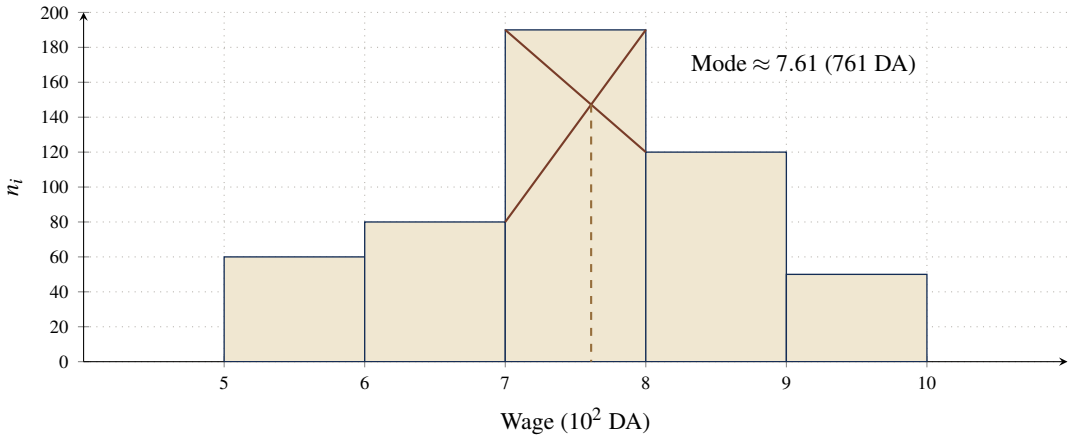


Figure 7: Histogram of the wages and graphical mode.

$$\text{Mode} \approx 7.61\ (10^2\ \text{DA}) = 761\ \text{DA}$$

which agrees with the calculated value 761.1 DA.

> **► Common Pitfall:**
> Joining the wrong corners (upper-*right* to upper-*right* of the neighbours instead of crossing them), or reading the mode on the vertical axis instead of the horizontal one.

> **✓ Sanity Check:**
> The crossing abscissa (7.61) lies inside $[7,8[$ and in its right half, as predicted by $120 > 80$ ✓.

4. Draw the polygon of cumulative ascending frequencies (ogive).

★★☆☆☆

Method *Ascending cumulative frequency polygon*

> ***Why we do this (Way of Thinking):***
> The ogive shows, for every wage x, the share of employees earning *at most* x. We know this share only at the class upper bounds, so we plot the points (upper bound, $F^{\nearrow}$) and join them with straight segments (the even-spread assumption). Before the first class nobody is below 5, so $F = 0$ up to $x = 5$, and the curve starts to rise only from there.

Execution: From the table, the points are $(5,0),(6,0.12),(7,0.28),(8,0.66),(9,0.90),(10,1)$, and the function is 0 on $[0,5]$.

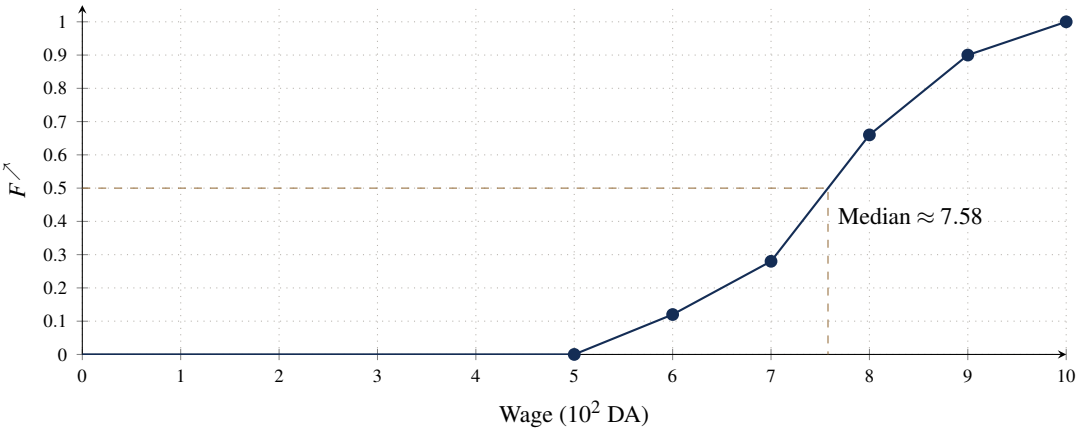


Figure 8: Ascending cumulative frequency polygon (ogive); the dashed line reads the median at height 0.5.

> **► Common Pitfall:**
> Plotting the cumulative values at the class centres or at the lower bounds, or starting the curve at a nonzero value. The ogive must be 0 before the first class and must end at 1 at the last upper bound.

> **✓ Sanity Check:**
> The curve is increasing and never decreases; it passes through $(5,0)$ and $(10,1)$. Reading it at height 0.5 gives $x \approx 7.58$, which will be confirmed by the calculation of the median in Question 5.

5. Calculate the mean salary. Is the distribution symmetrical? Justify your answer. Compute the variance and standard deviation. ★★★★☆

Method *Mean from class centres; mode–median–mean comparison; variance by the shortcut formula*

> ***Why we do this (Way of Thinking):***
> For a symmetric distribution the three "typical values" (mode, median, mean) sit at the same place, so we compare them: that is the justification the question asks for. The mode is already known (7.611); we compute the median by interpolation and the mean from the class centres. For the variance we use the shortcut "mean of the squares minus the square of the mean", which avoids computing a deviation for every class.

Execution: Mean. From the table, $\sum n_i c_i = 330+520+1425+1020+475$: $330+520=850$, $850+1425=2275$, $2275+1020=3295$, $3295+475=3770$. Hence

$$\bar{x} = \frac{1}{N}\sum_i n_i c_i = \frac{3770}{500} = 7.54 \ (\times 10^2 \text{ DA})$$

Median. $\dfrac{N}{2} = 250$ and $140 < 250 \leq 330$, so the median class is $[7,8[$ (previous ascending frequency 0.28, class frequency 0.38):

$$\text{Median} = 7 + 1 \times \frac{0.5-0.28}{0.38} = 7 + \frac{0.22}{0.38} = 7 + 0.5789 \approx 7.579 \ (\times 10^2 \text{ DA}).$$

Symmetry. In DA: Mode $= 761.1$, Median ≈ 757.9, $\bar{x} = 754$. These three values are very close (the gaps are less than 8 DA, against a standard deviation of about 113 DA, computed below). The distribution is therefore *approximately symmetric*, with a very slight skew to the left (since Mode $>$ Median $> \bar{x}$).

Variance. Squares of the centres: $5.5^2 = 30.25$, $6.5^2 = 42.25$, $7.5^2 = 56.25$, $8.5^2 = 72.25$, $9.5^2 = 90.25$. Then $\sum n_i c_i^2 = 60\times 30.25 + 80\times 42.25 + 190\times 56.25 + 120\times 72.25 + 50\times 90.25 = 1815 + 3380 + 10687.5 + 8670 + 4512.5 = 29065$.

$$V(X) = \frac{1}{N}\sum_i n_i c_i^2 - \bar{x}^2 = \frac{29065}{500} - 7.54^2 = 58.13 - 56.8516 = 1.2784 \ (\times 10^2 \text{ DA})^2$$

Converting to DA: $V(X) = 1.2784 \times 100^2 = 12\,784$ DA^2. Standard deviation: $\sigma(X) = \sqrt{1.2784} \approx 1.1307$, i.e. $1.1307 \times 100 \approx 113.07$ DA.

$$\bar{x} = 754 \text{ DA}, \quad V(X) \approx 12\,784 \text{ DA}^2, \quad \sigma(X) \approx 113.07 \text{ DA}$$

► Common Pitfall:
Forgetting to subtract $\bar{x}^2$ (or subtracting $\bar{x}$ instead of $\bar{x}^2$) in the variance; mixing units (the variance is in $(10^2 \text{ DA})^2$ and must be multiplied by 100^2, while σ is multiplied by 100); and claiming symmetry from the mean alone instead of comparing mode, median and mean.

✓ Sanity Check:
A variance is never negative; the mean 7.54 lies between the smallest and largest centres (5.5 and 9.5); cross-check with deviations from the mean: $\frac{1}{500}\left[60(2.04)^2 + 80(1.04)^2 + 190(0.04)^2 + 120(0.96)^2 + 50(1.96)^2\right] = \frac{639.2}{500} = 1.2784$ ✓.

6. Estimate the number of employees with salaries between 640 DA and 850 DA. ★★★☆☆

Method *Estimation by linear interpolation inside classes (two equivalent methods)*

Why we do this (Way of Thinking):
First convert to the table's unit: 640 DA $= 6.4$ and 850 DA $= 8.5$ (in 10^2 DA). These two values fall in the middle of the classes $[6,7[$ and $[8,9[$, and the class $[7,8[$ is entirely inside. As in Exercise 2, we assume an even spread of wages inside each class and keep the proportional part. The CDF method does the same in one line: the share between two wages is $F(8.5) - F(6.4)$.

Execution: Method 1 (proportion of each class).

$$\begin{aligned} [6.4,7[&: \ 80 \times \frac{7-6.4}{1} = 80 \times 0.6 = 48, \\ [7,8[&: \ 190 \text{ (whole class)}, \\ [8,8.5[&: \ 120 \times \frac{8.5-8}{1} = 120 \times 0.5 = 60. \end{aligned}$$

$$n = 48 + 190 + 60 = 298$$

Method 2 (using the CDF). With $F(x) = F(a) + \dfrac{x-a}{b-a}\big(F(b) - F(a)\big)$ for $a \le x \le b$ (linear between class bounds), and $n = N\big(F(x_2) - F(x_1)\big)$:

$$F(6.4) = F(6) + \frac{6.4-6}{7-6}\big(F(7) - F(6)\big) = 0.12 + 0.4\,(0.28 - 0.12) = 0.12 + 0.064 = 0.184,$$
$$F(8.5) = F(8) + \frac{8.5-8}{9-8}\big(F(9) - F(8)\big) = 0.66 + 0.5\,(0.90 - 0.66) = 0.66 + 0.12 = 0.78.$$

$$n = 500 \times \big(F(8.5) - F(6.4)\big) = 500 \times (0.78 - 0.184) = 500 \times 0.596 = 298$$

$$n = 298 \text{ employees}$$

► Common Pitfall:
Forgetting the conversion 640 $\to$ 6.4 and 850 $\to$ 8.5, or using the part $7 - 6.4$ of the class $[6,7[$ on the wrong side: the employees we want are those *above* 6.4, i.e. the part from 6.4 to 7 (0.6 of the class), not from 6 to 6.4.

> **✓ Sanity Check:**
> The estimate must lie between the count of the class fully inside (190) and the counts of all classes touched ($80+190+120=390$): $190 < 298 < 390$ ✓. Both methods give 298.

7. If each employee receives a flat raise of 100 DA plus a 5 % bonus on their initial salary, what are the new mean and standard deviation? ★★★☆☆

Method ***Affine transformation $Y = aX + b$ of a statistical variable***

> ***Why we do this (Way of Thinking):***
> The new wage of each employee is built from his old one: he gets his old wage X, plus 100 DA, plus 5 % of X, i.e. $0.05X$. So $Y = X + 100 + 0.05X = 1.05X + 100$, which is an *affine* transformation $aX + b$ with $a = 1.05$ and $b = 100$. Such a transformation has simple effects: the mean undergoes the same transformation, $\bar{y} = a\bar{x} + b$, but the standard deviation ignores the shift b (adding a constant to everybody does not change how spread out the wages are) and is only scaled by $|a|$.

Execution:

Sub **Let X be the initial wage (in DA) and Y the new wage: $Y = X + 100 + 0.05X = 1.05X + 100$, so $a = 1.05$, $b = 100$.**

Mean (with $\bar{x} = 754$ DA):

$$\bar{y} = 1.05 \times 754 + 100 = 791.7 + 100 = 891.7 \text{ DA}.$$

Variance and standard deviation (with $V(X) \approx 12\,784$ DA^2 and $\sigma(X) \approx 113.07$ DA):

$$V(Y) = a^2 V(X) = 1.05^2 \times 12\,784 = 1.1025 \times 12\,784 \approx 14\,094 \text{ DA}^2,$$
$$\sigma(Y) = |a|\,\sigma(X) = 1.05 \times 113.07 \approx 118.72 \text{ DA}.$$

$$\boxed{\bar{y} = 891.7 \text{ DA}, \qquad \sigma(Y) \approx 118.72 \text{ DA}}$$

> **► Common Pitfall:**
> Writing $Y = X + 100 + 5$ (the 5 % is $0.05X$, not 5), multiplying the standard deviation by a^2 (that factor belongs to the variance), or adding 100 to $\sigma(X)$.

> **✓ Sanity Check:**
> Mean: $754 + 100 + 0.05 \times 754 = 754 + 100 + 37.7 = 891.7$ ✓. Standard deviation: $\sqrt{14\,094.36} \approx 118.72 = 1.05 \times 113.07$ ✓. The new wages are more spread out (σ increases) because the bonus is proportional to the wage.