

Data visualization

Visual encoding, accessible charts and honest claims

Mohammad Sabokrou | New Uzbekistan University

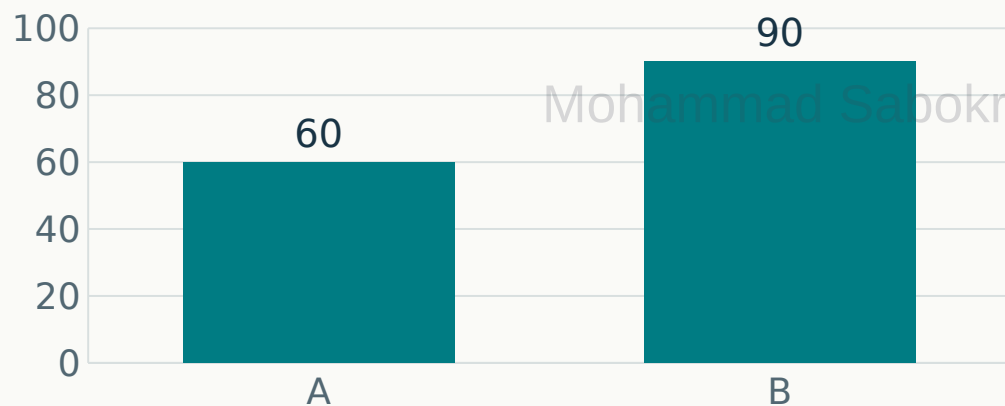
What does the chart make us see?

Mohammad Sabokrou

Bar baselines change perceived differences

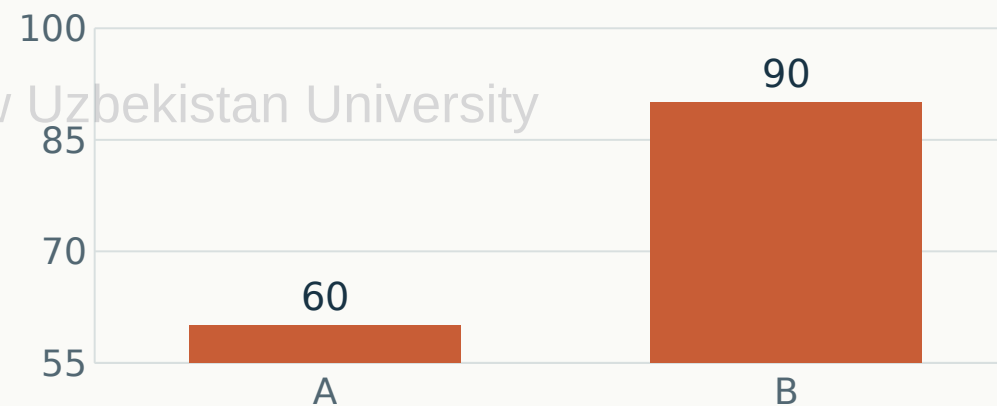
Same values, 60 and 90. The second chart starts at 55.

Baseline 0: lengths 60 and 90



Actual ratio: $90 \div 60 = 1.5$

Baseline 55: lengths 5 and 35



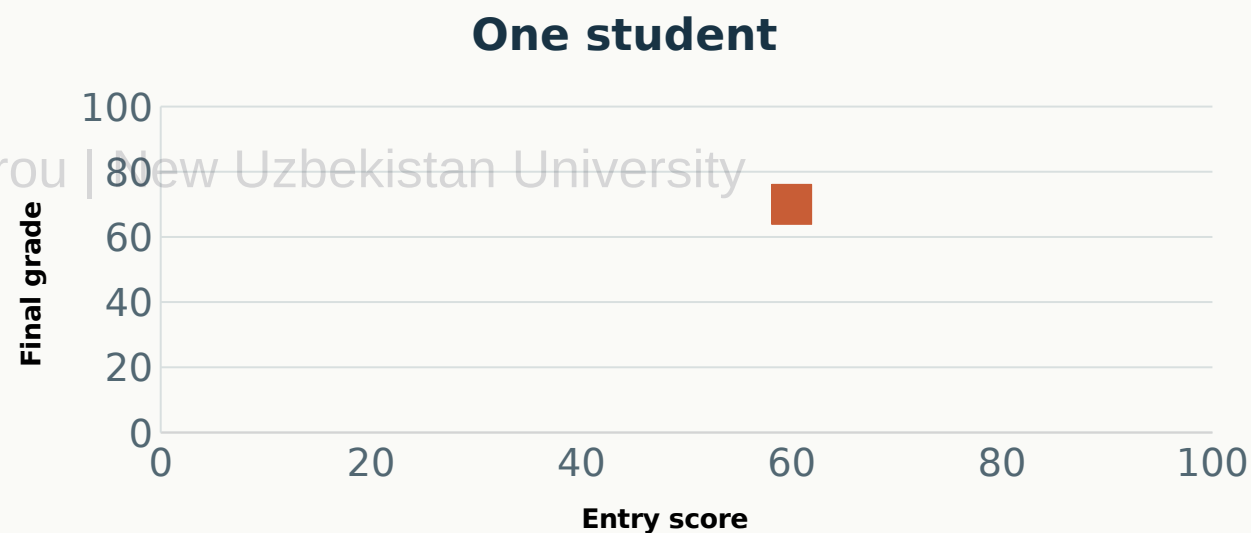
Visible ratio: $(90 - 55) \div (60 - 55) = 7$

90 is 50% higher than 60, although the truncated bar looks 7 times as long

Visual encoding: a row becomes a mark

A mark is a point, bar or line. Position, length, colour and area are visual channels.

Field	Value
Entry score	60
Final grade	70
Study mode	Part-time

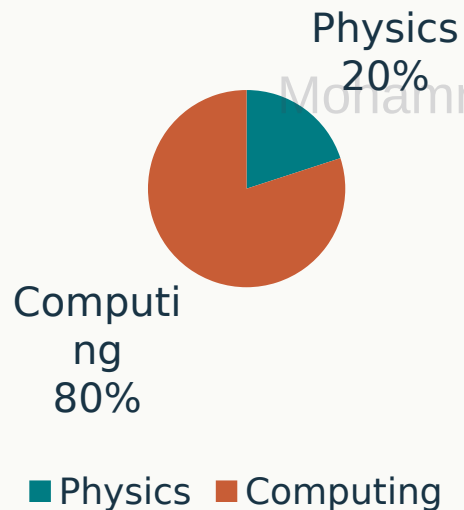


This square represents one student: x = entry score, y = final grade, shape = study mode.

A pie chart divides one total into shares

Illustrative subjects: Physics 100, Computing 400. Total = 500.

Students by subject



$$\text{Physics: } 100 \div 500 = 20\%$$

$$\text{Computing: } 400 \div 500 = 80\%$$

$$\begin{aligned}\text{Slice angles: } 0.20 \times 360^\circ &= 72^\circ \\ 0.80 \times 360^\circ &= 288^\circ\end{aligned}$$

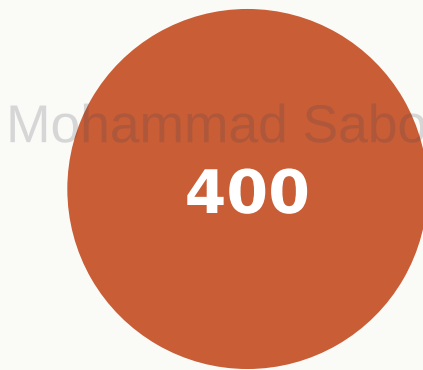
Pie slices share one radius. Their angles and areas represent shares of the total.

Bubble area and radius

The two labelled circles encode student counts through their area. A = area and r = radius.



Physics: 100



Computing: 400

Circle area: $A = \pi r^2$

Area ratio = $400 \div 100 = 4$

Radius ratio = $\sqrt{4} = 2$

Example radii: 10 and 20

4 times the value needs 4 times the area, so the radius grows only 2 times

Chart choice follows the question

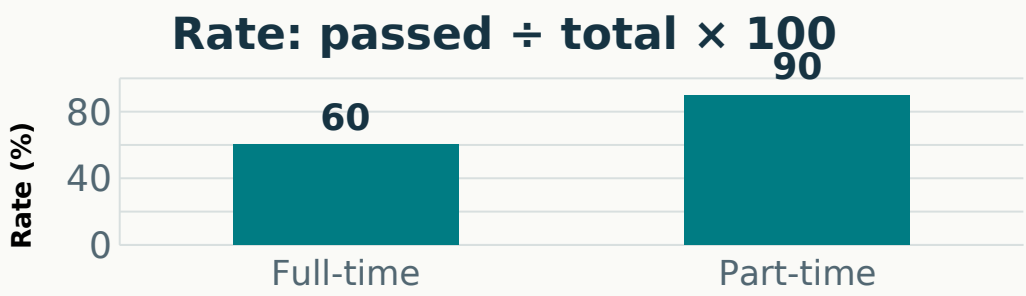
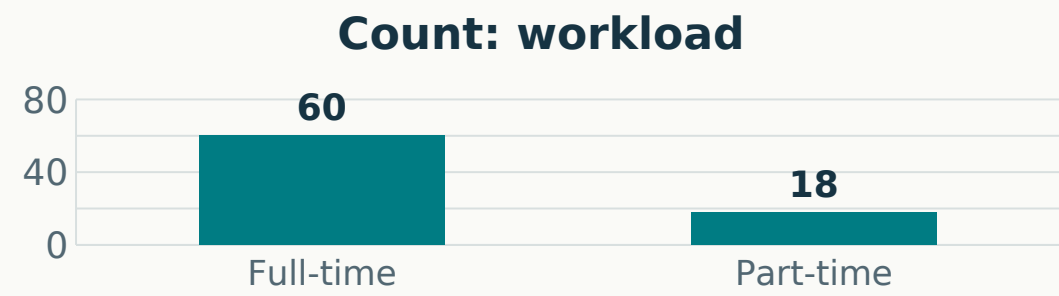
Variable type narrows the options. The comparison selects the design.

Question	Variables	Useful starting point
How many in each category?	Categorical + count	Bar chart
How are grades distributed?	One numerical variable	Histogram or box plot
Do two scores vary together?	Two numerical variables	Scatter plot
How does activity change?	Time + numerical variable	Line chart
How do groups differ?	Group + numerical variable	Grouped plot or small multiples

Counts and rates answer different questions

Constructed registrations. Full-time: 60 passes out of 100. Part-time: 18 out of 20.

Study mode	Passed	Did not pass	All registrations
Full-time	60	40	100
Part-time	18	2	20



Full-time has more passes. Part-time has the higher pass rate: $60/100 = 60\%$, $18/20 = 90\%$.

Percentage points and relative change

Illustrative pass rates increase from 60% to 90%.

**Absolute
difference**

$$90\% - 60\% = 30 \text{ percentage points}$$

**Relative
increase**

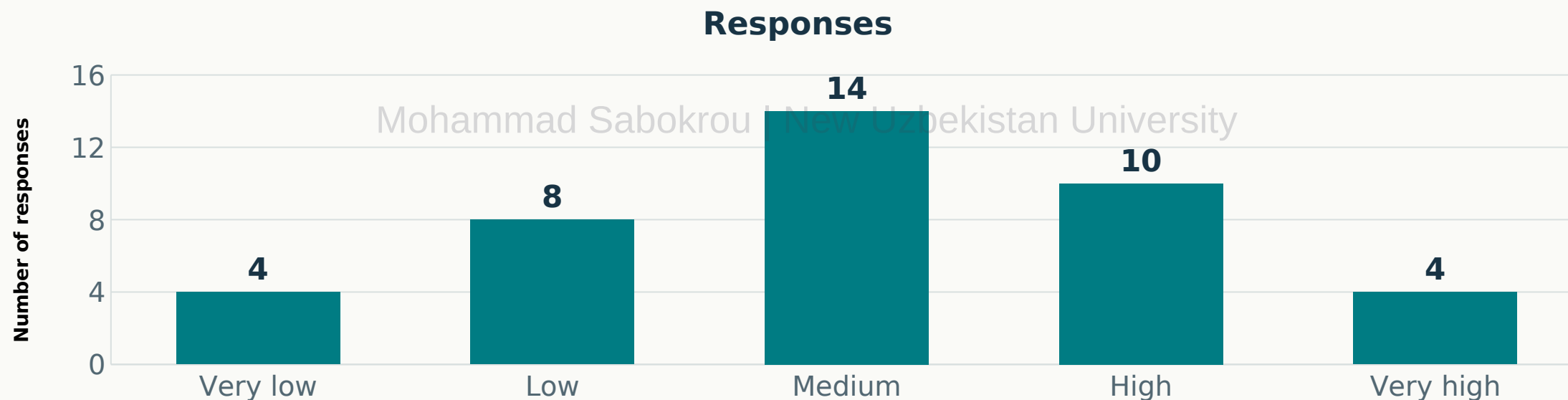
$$(90 - 60) / 60 = 50\%$$

**Defensible
wording**

“30 percentage points higher” or “50% higher relative to 60%”

Category order carries meaning

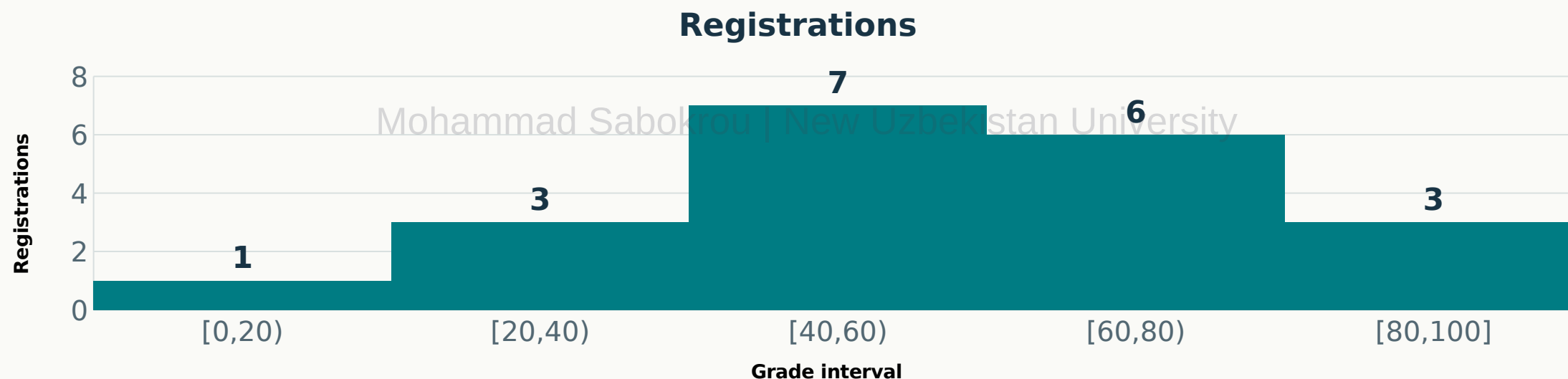
Illustrative survey responses retain their natural order.



Ordered categories keep their sequence. Unordered categories may suit ranking.

Histograms group numerical observations

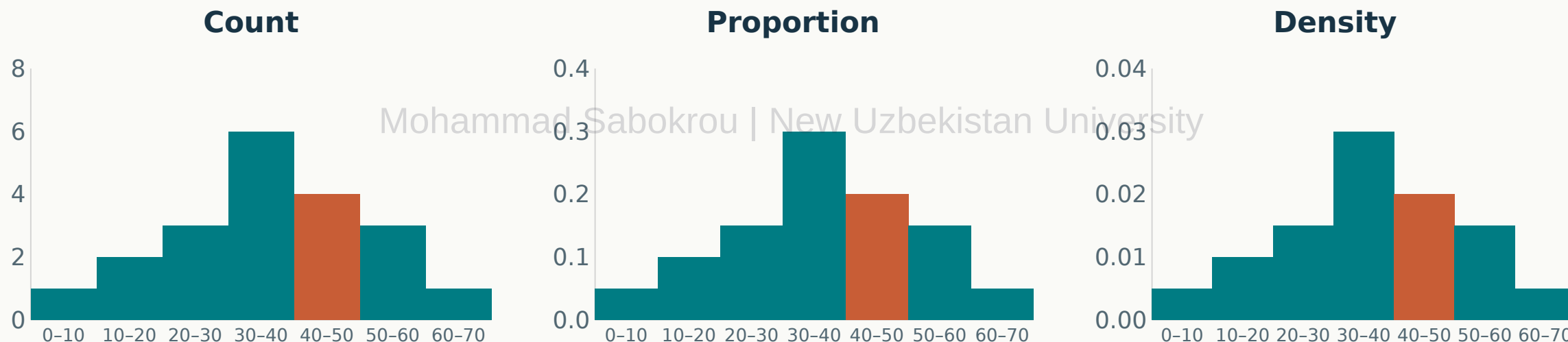
Touching bins cover adjacent numerical intervals. Each grade enters exactly one bin.



[40,60) includes 40 and excludes 60. Category bars, unlike histogram bins, keep gaps.

The same histogram with three vertical axes

20 observations. Bin counts = 1, 2, 3, 6, 4, 3, 1. Each width = 10.

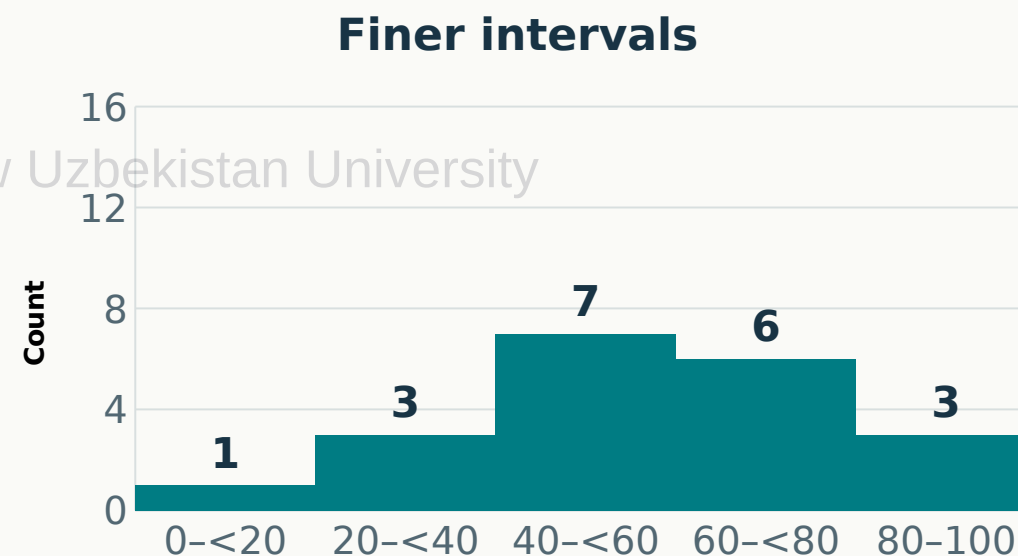
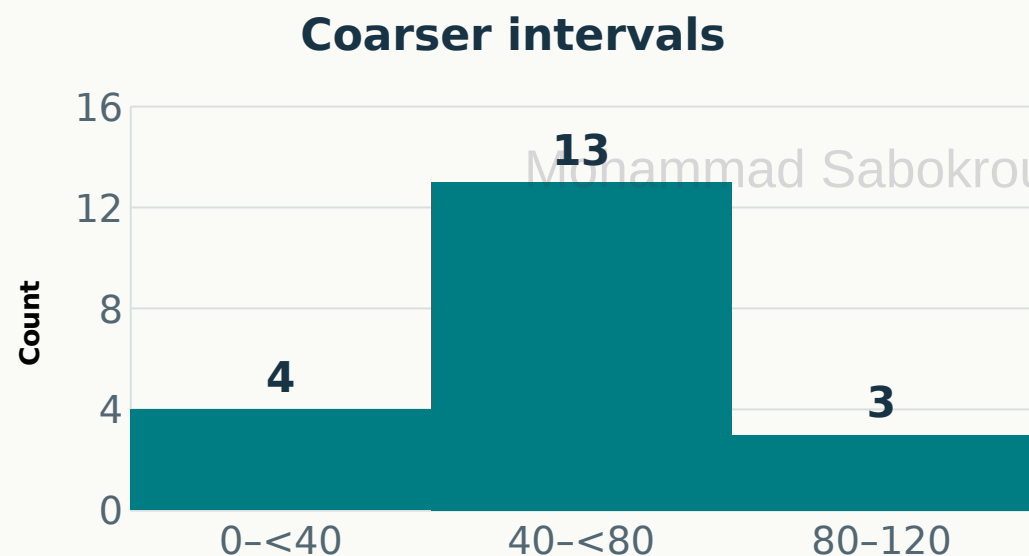


Highlighted bin 40-50: $c = 4$, $n = 20$, width = 10

Count = 4. Proportion = $4/20 = 0.20$. Density = $4/(20 \times 10) = 0.020$.

Bin width changes visible detail

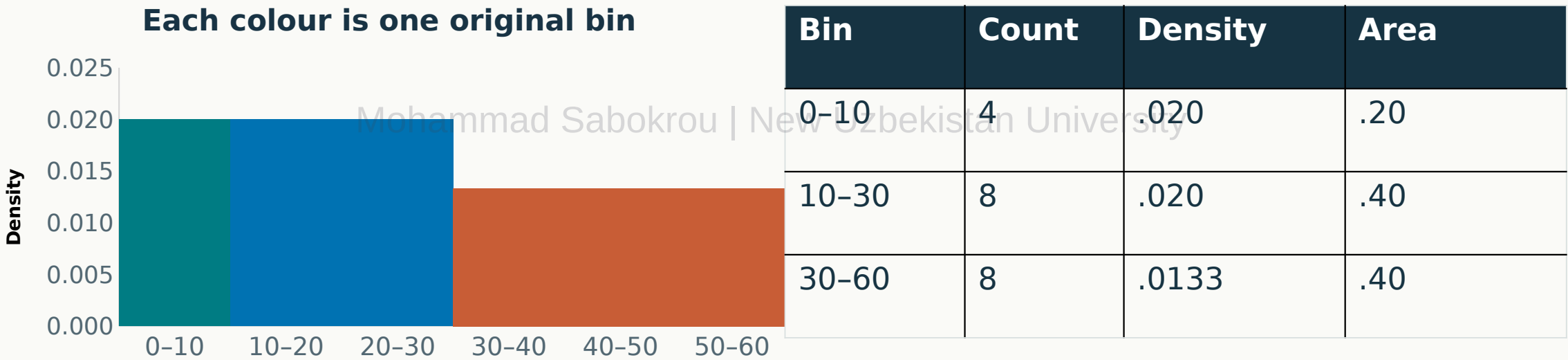
The same 20 grades, grouped in two different ways.



Both totals are 20. A pattern should survive reasonable bin choices.

Unequal widths: height and amount differ

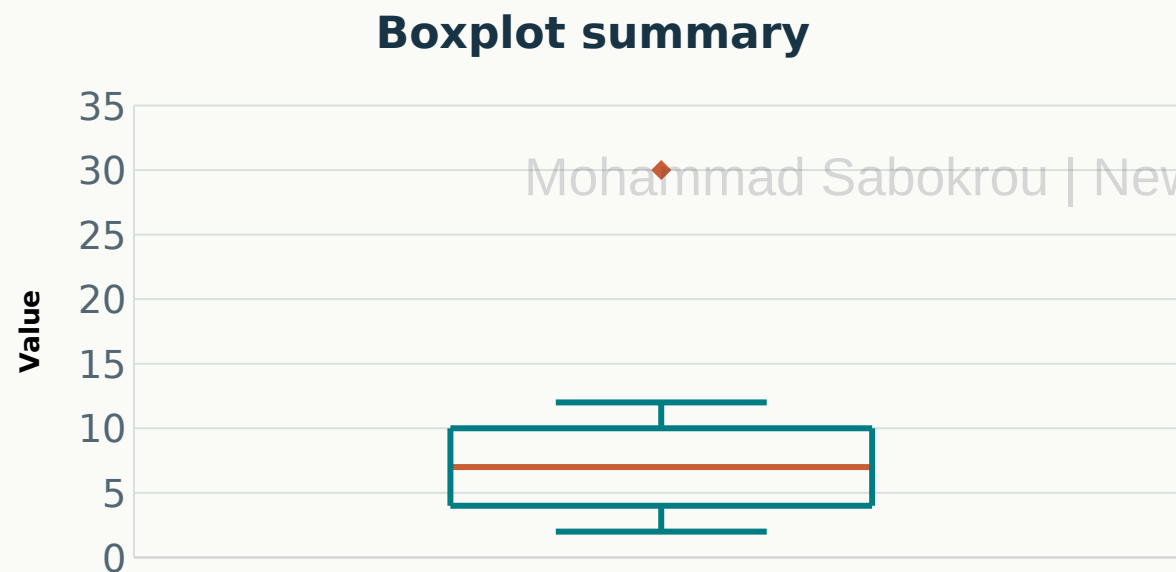
Constructed sample of 20 observations. Original bins have widths 10, 20 and 30.



The first two bins have equal density. The wider bin contains twice as many observations.

A box plot summarizes location and spread

Illustrative values: 2, 4, 4, 4, 6, 8, 10, 10, 12, 30



Box

Q1 = 4 to Q3 = 10

Middle line

Median = 7

Flag

Diamond at 30:
beyond the upper
fence

$$\text{IQR} = \text{Q3} - \text{Q1} = 10 - 4 = 6$$

Fences and whiskers are different

For $Q1=4$, $Q3=10$ and $IQR=6$ in the same sample.

Lower fence $Q1 - 1.5 \times IQR = 4 - 9 = -5$

Upper fence $Q3 + 1.5 \times IQR = 10 + 9 = 19$

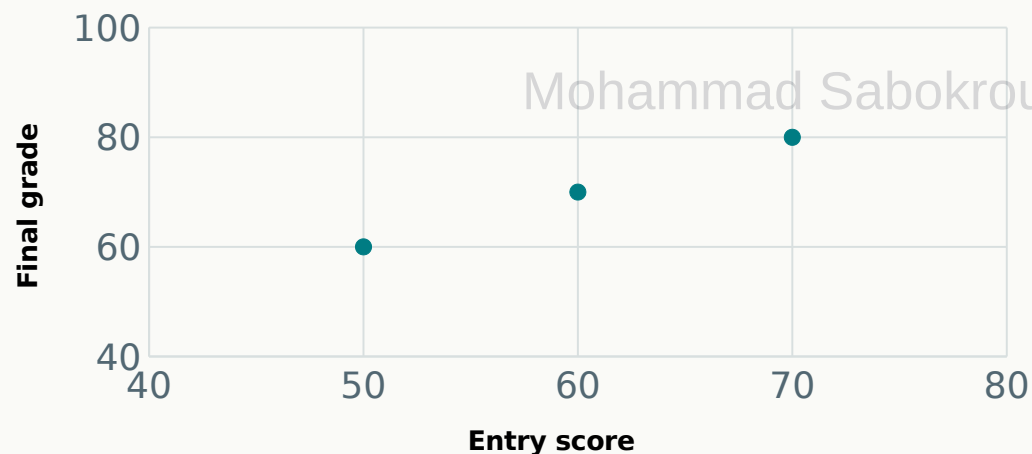
Observed whiskers Smallest and largest values inside the fences: 2 and 12

30 receives an outlier flag. A flag does not prove a data error.

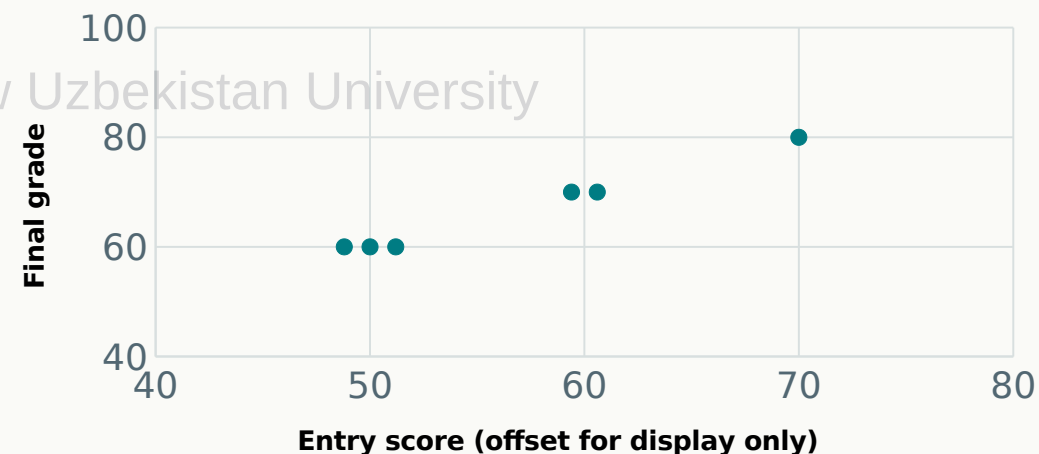
Overlapping observations and display jitter

Six students: three at (50,60), two at (60,70), and one at (70,80).

Original: 3 visible positions



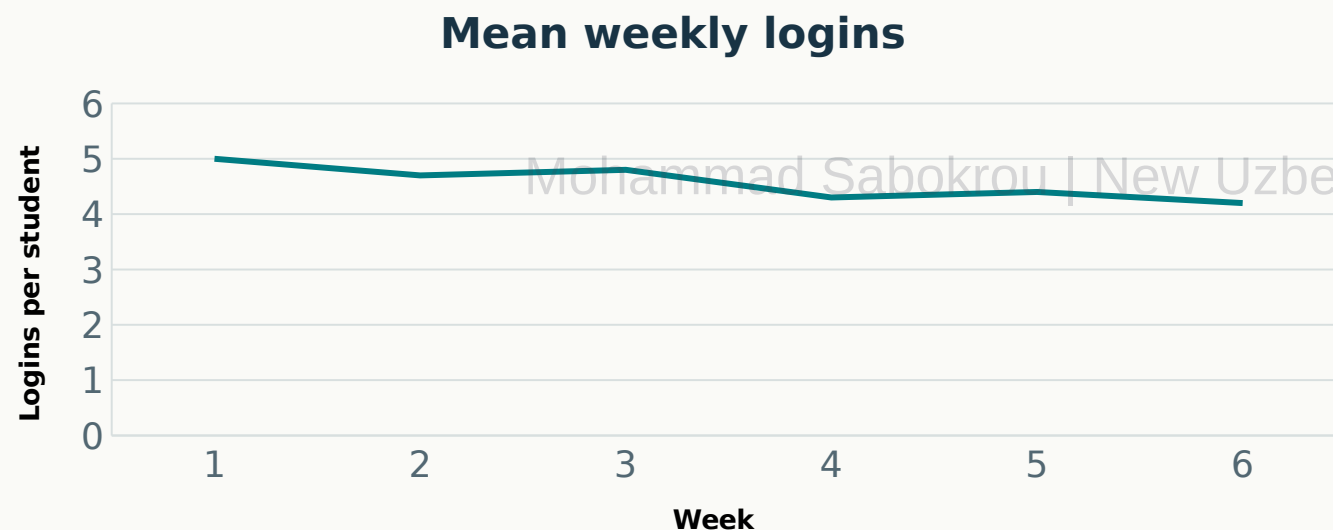
Jitter: all 6 records visible



Jitter uses small disclosed offsets. It reveals duplicate pairs while preserving all six records.

A line connects observations in time order

Fixed cohort of 10 students. A weekly mean divides total logins by 10.



$$\text{Week 1: } 50 \div 10 = 5$$

$$\text{Week 6: } 42 \div 10 = 4.2$$

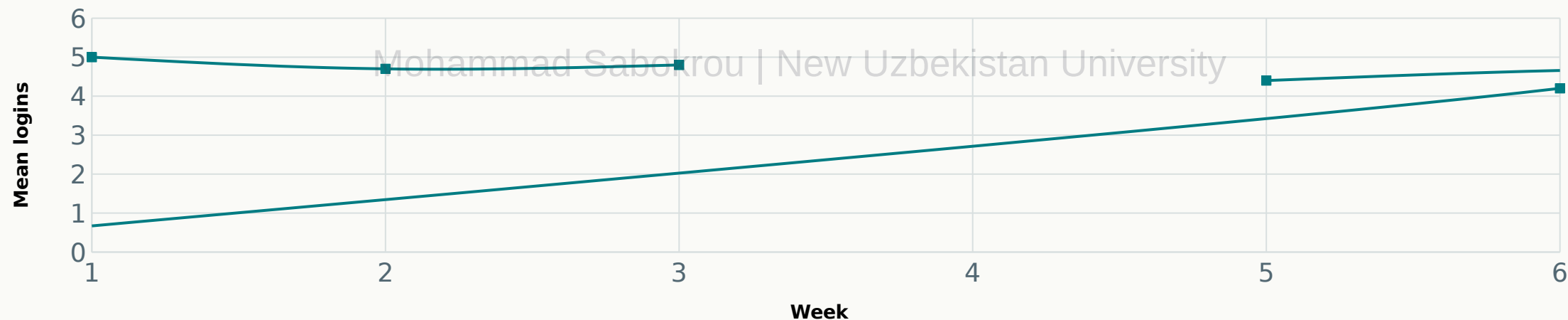
$$\text{Change: } 4.2 - 5 = -0.8$$

Equal horizontal steps mean equal time intervals. The line shows a general decline.

Missing observations are different from zero

Week 4 has no recorded value in this illustrative series.

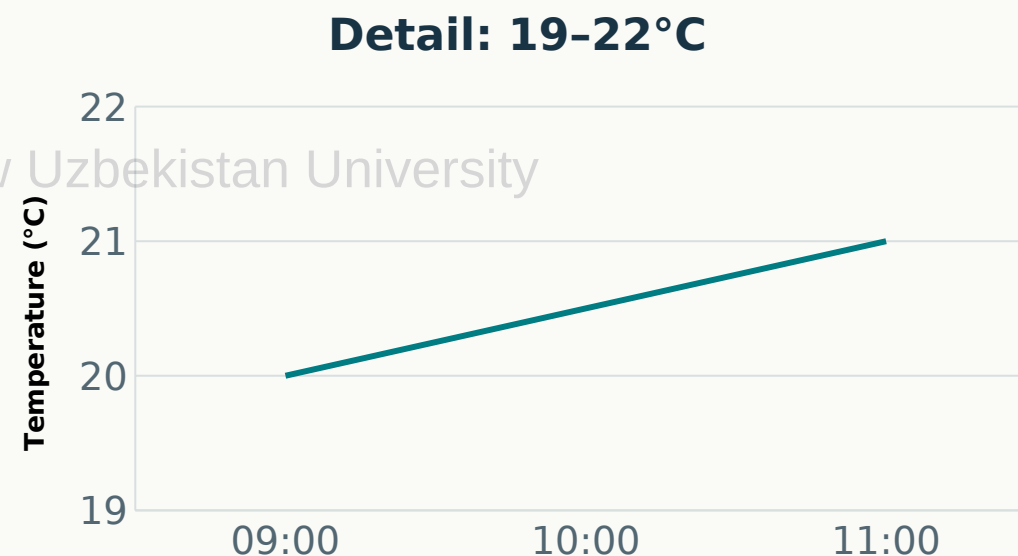
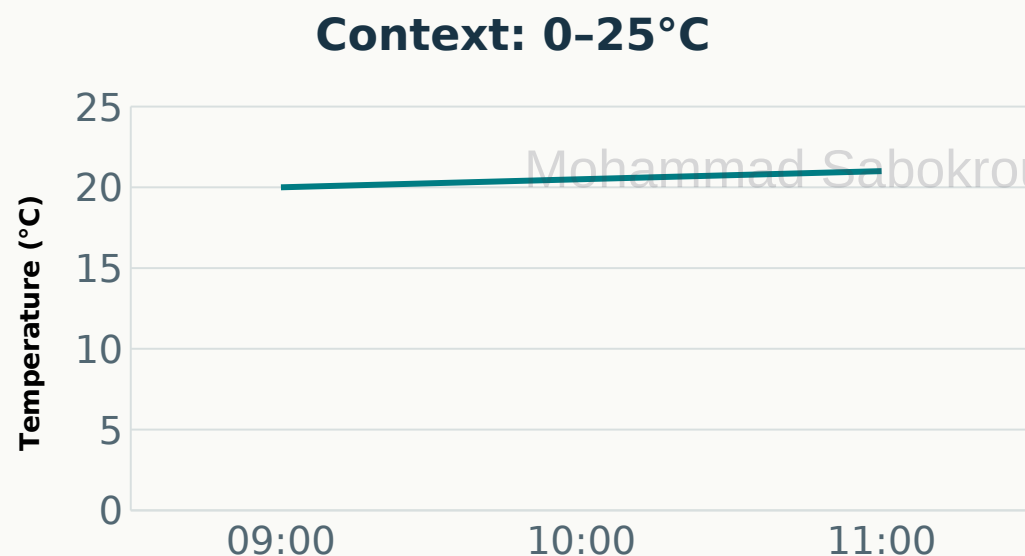
Observed segments, with a gap at week 4



A gap means “unobserved”. A zero means an observed value of zero.

A line chart can focus on a small change

The same temperatures appear in both panels.



Both show a 1°C increase. A focused line axis can reveal detail when its limits remain explicit.

10-minute break

When we return: scales, accessibility and chart critique

One question to keep in mind

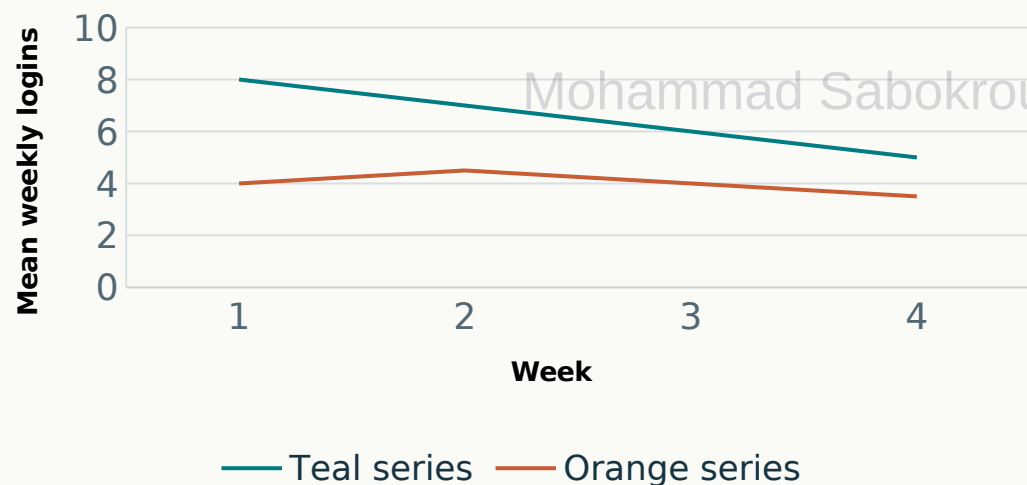
Mohammad Sabokrou | New Uzbekistan University

Would the message survive without its colours?

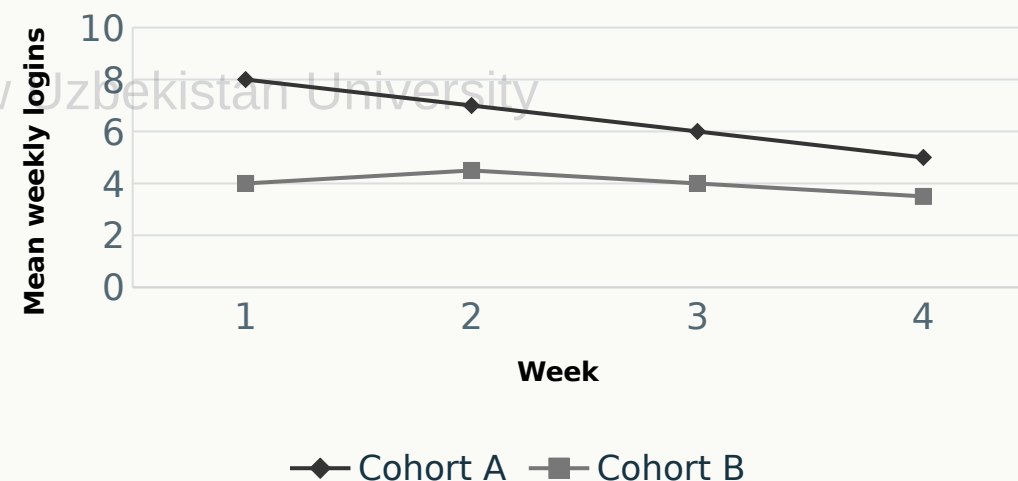
Group identity with and without colour

The same data in both panels. Names and marker shapes support reading in grayscale.

Identity depends on colour



Names and shapes in grayscale

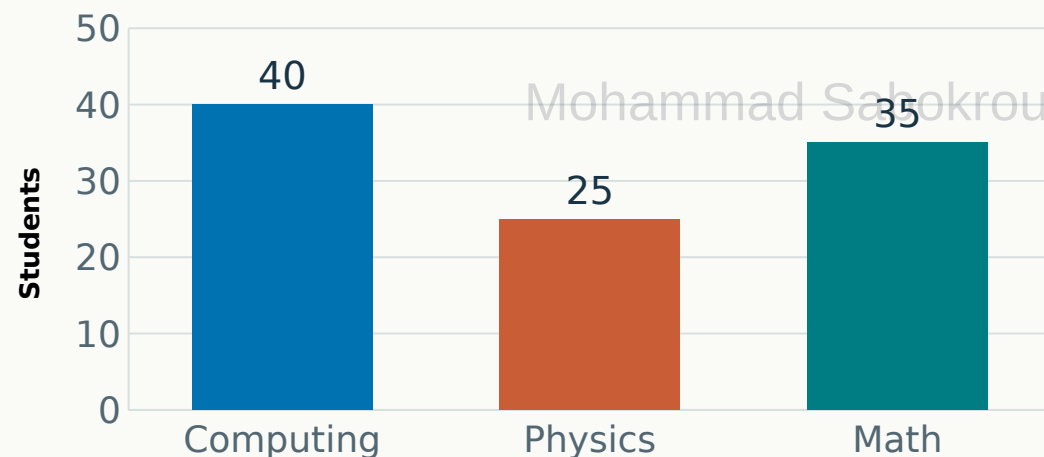


Check the chart in grayscale and at classroom size. Keep labels, type and contrast readable.

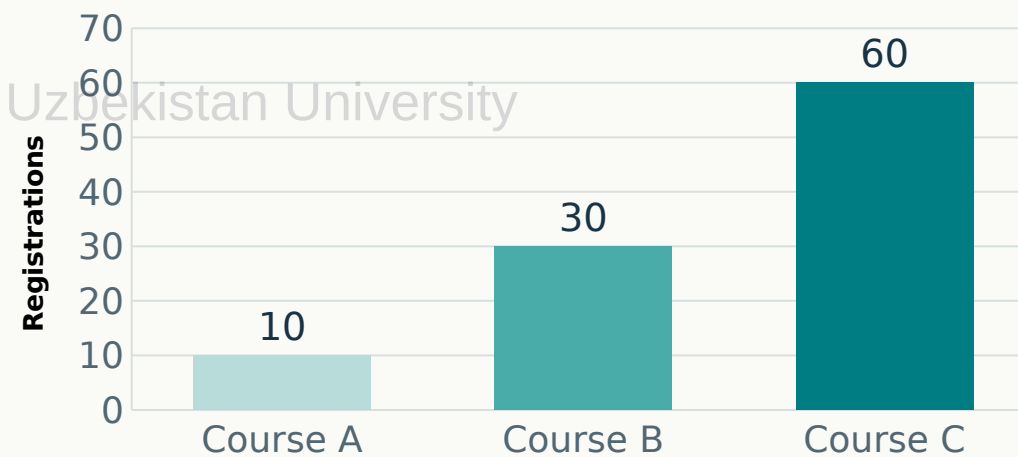
Colour distinguishes categories or orders values

Illustrative examples. A label always accompanies colour.

Categories: separate identities



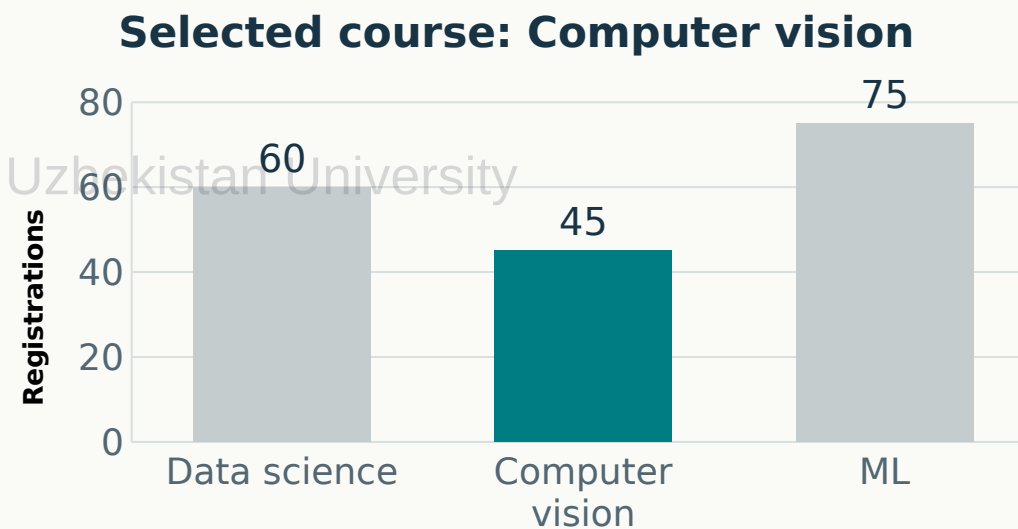
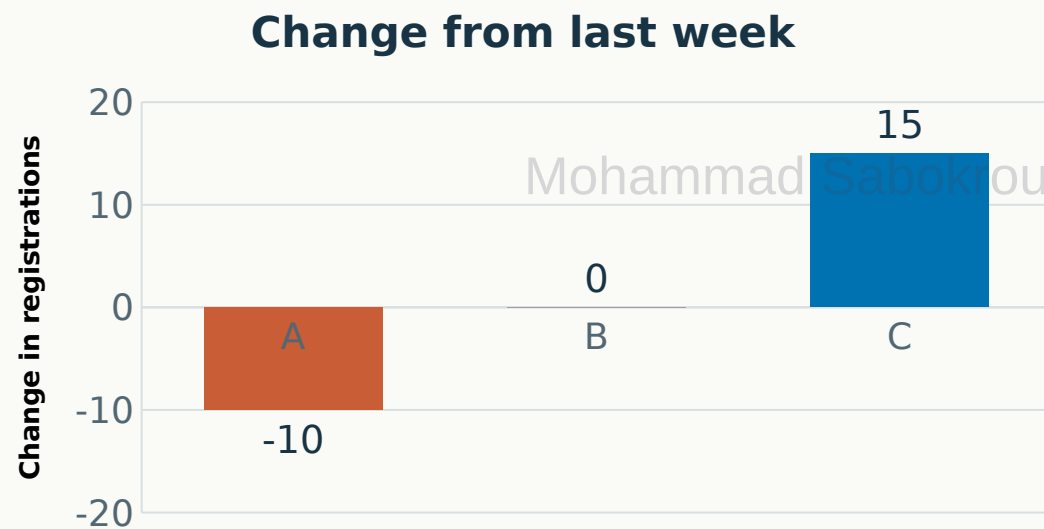
Ordered values: pale to strong



Different hues identify subjects. Increasing colour intensity suggests increasing quantity.

Colour shows direction or focuses attention

Two further strategies: deviation around zero and one selected group.

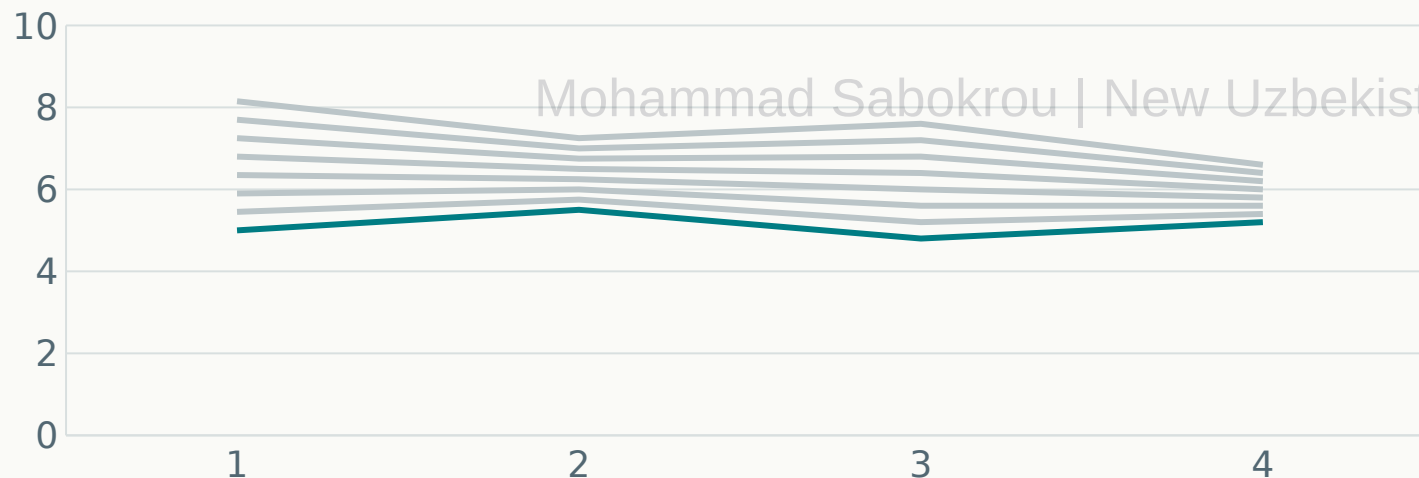


Zero separates decreases and increases. One highlight makes the selected course easy to find.

Highlighting one trajectory

Question: did course C1 activity increase from week 1 to week 4?

C1 highlighted, other courses provide context



C1: 5.0 to 5.2

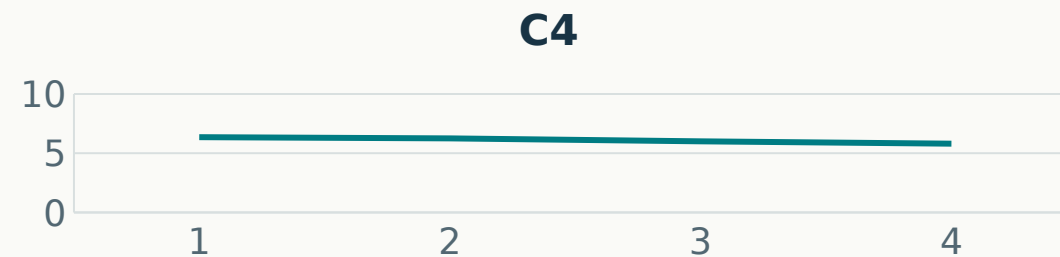
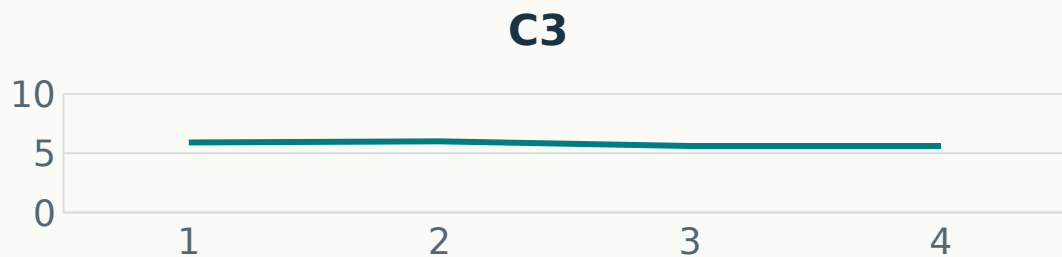
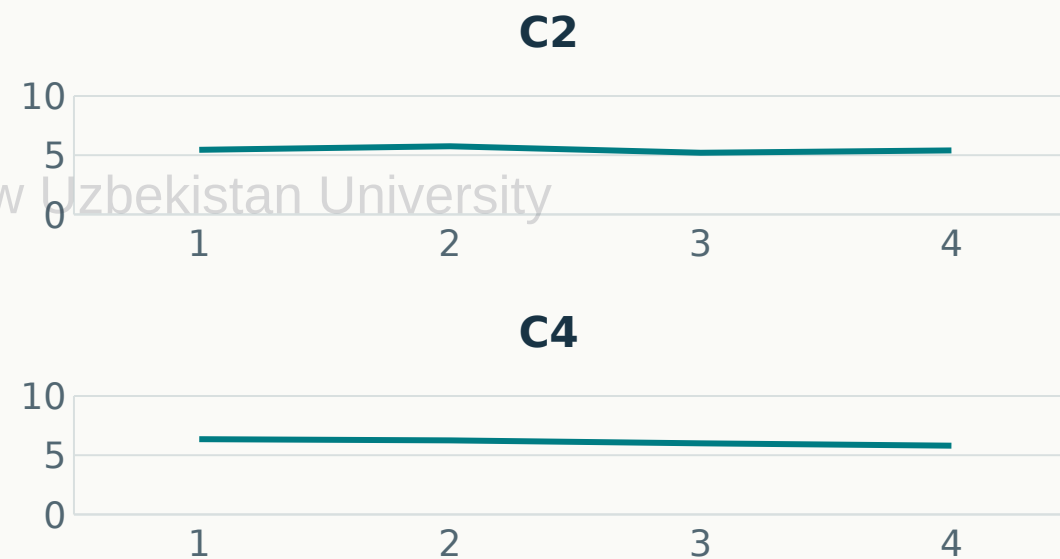
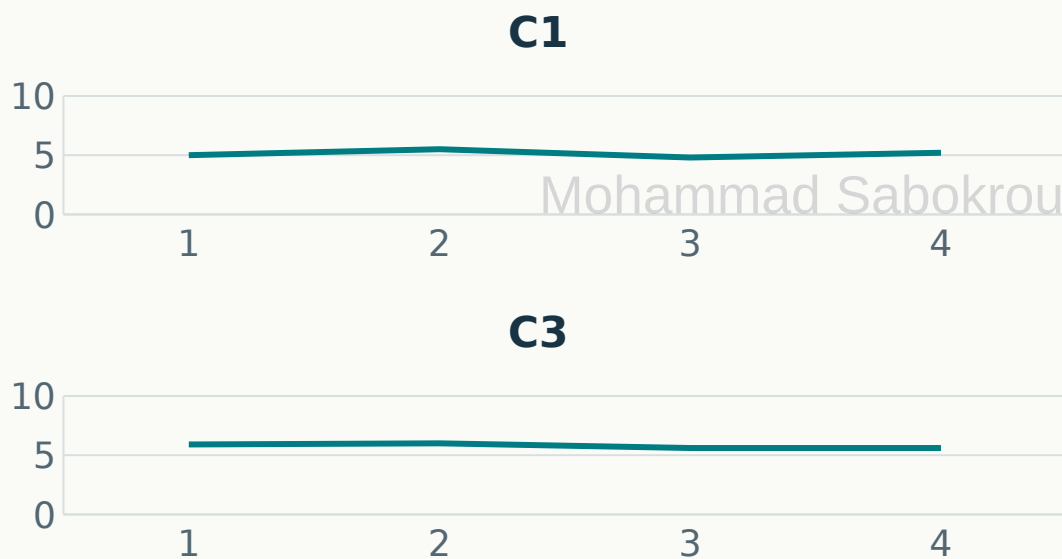
$$\text{Change} = 5.2 - 5.0 = +0.2$$

$$\begin{aligned}\text{Relative change} &= 0.2/5 \times 100 \\ &= +4\%\end{aligned}$$

One question, one highlighted series. The other seven courses stay visible.

Small multiples on a common scale

C1-C4 from the previous chart. Every panel shares the vertical range 0-10.



At week 4: C1 = 5.2, C4 = 5.8. Difference = $5.8 - 5.2 = 0.6$.

Pooled rates are weighted by group sizes

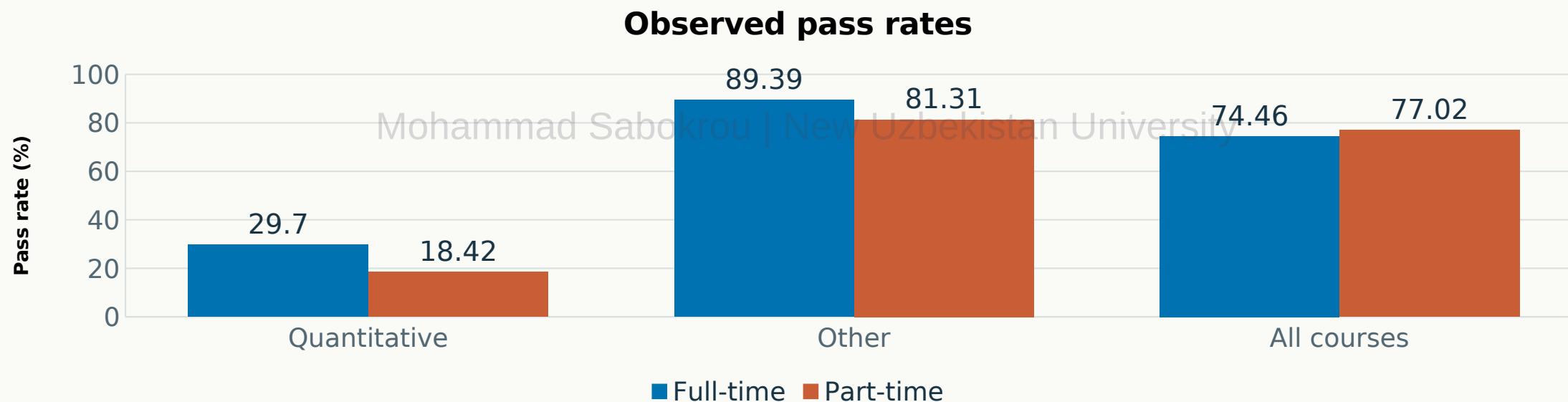
Synthetic registrations from Week 5. A registration can differ from a distinct student.

Course group	Full-time passes / all	Part-time passes / all
Quantitative	$305/1,027 = 29.7\%$	$7/38 = 18.4\%$
Other	$2,754/3,081 = 89.4\%$	$422/519 = 81.3\%$
Pooled	$(305+2,754)/(1,027+3,081)$ $= 3,059/4,108 = 74.5\%$	$(7+422)/(38+519)$ $= 429/557 = 77.0\%$

Pooling adds counts before dividing. It does not take the simple mean of the percentages.

A grouped chart reveals the reversal

Same counts as the previous slide. “All courses” pools the other two groups.

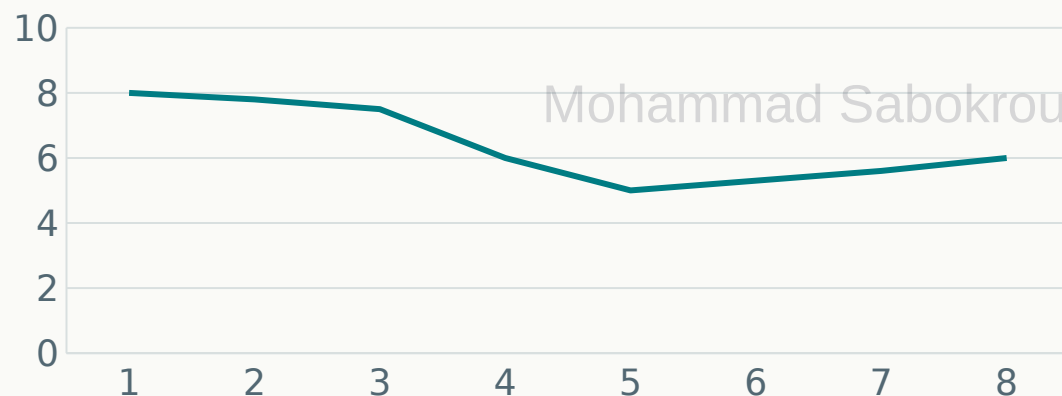


Full-time leads within each group. The pooled ordering reverses. These observations do not show cause.

The time window changes the comparison

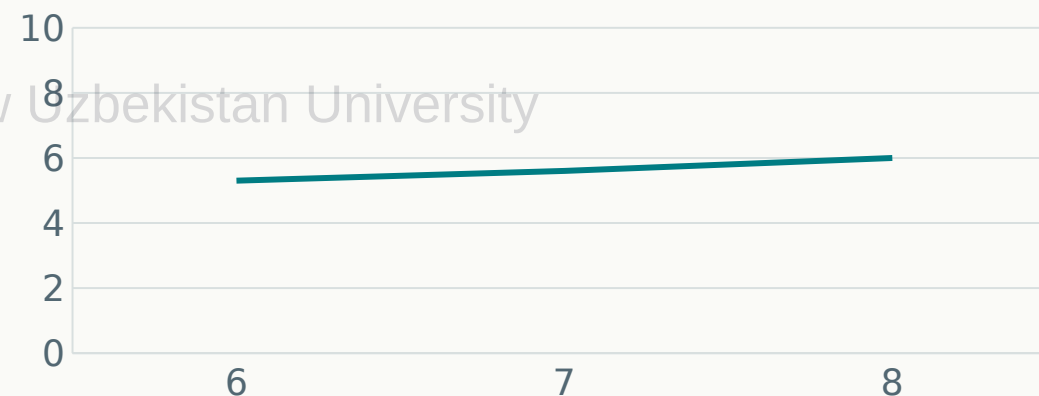
Fixed cohort. Same 0–10 scale in both panels.

Whole period: weeks 1–8



Whole period: $(6 - 8)/8 \times 100 = -25\%$

Recent period: weeks 6–8



Recent period: $(6 - 5.3)/5.3 \times 100 \approx +13.2\%$

Recent recovery and a full-period decline are both true. The headline needs dates.

A dashboard flags a change for investigation

Participation = active ÷ eligible × 100. Active means at least one login in the stated week.

Course	Previous week	Current week	Change
C1	82/100 = 82%	80/100 = 80%	−2 points
C2	30/40 = 75%	24/40 = 60%	−15 points
C3	18/20 = 90%	18/20 = 90%	0 points

C2: 60% − 75% = −15 percentage points
Relative change = (60 − 75)/75 × 100 = −20%

An aggregate can still reveal a person

Constructed example. No actual student information appears here.

Published result	What someone could infer
Four students in a group. Three passed.	Individual results remain unknown from this count alone.
A second filter shows three students. All three passed.	The one excluded student did not pass.

Inference: failures in full group = $4 - 3 = 1$
Failures in filtered group = $3 - 3 = 0$
The excluded person accounts for the remaining failure.

Design task: rework volume and defect rate

Constructed inspections. Choose a chart for each question and justify the comparison.

Production line	Inspected items	Defective items
A	200	20
B	50	10

- 1. Which line has more items to rework? Draw a chart of the counts.
- 2. Which line has the higher defect rate? Calculate and plot both rates.

Label the denominators and axes. Write one claim supported by these observations.

Chart review checklist

Apply these checks to the factory charts you have just designed.

Check	Question	Answer for the example
Unit	What does one record mean?	One inspected item
Calculation	Where does 20% come from?	Line B: $10/50 \times 100$
Scale	Where do the bars start?	0, on a clearly labelled common axis
Identity	How can readers identify groups?	Production-line names label the bars
Claim	Does this prove a causal effect?	No. Observed comparison only

A useful critique points to actual numbers, scales and labels in the figure.

Exit questions and preparation

Week 7 is revision. Week 8 is the midterm examination.

Self-check	Explain your reasoning
A rate rises from 20% to 30%	Find the percentage-point difference and the relative increase.
A density bin has height .02 and width 10	What fraction of the observations lies in this bin?
A bubble represents four times the value	How should its area and radius change?

Reading: OpenIntro Statistics §§2.1-2.2. ISLP Python edition §2.3.

Optional reference: spread and uncertainty

Preview of later inference. $n = 25$, sample mean = 70, sample standard deviation = 10.

Quantity	Calculation	Meaning
Sample SD	$s = 10$ grade points	Spread among individual grades
Standard error	$SE = s/\sqrt{n} = 10/\sqrt{25} = 2$	Estimated uncertainty of the mean
Approx. 95% interval	$70 \pm 2.064 \times 2 = [65.87, 74.13]$	t interval for the population mean

Every interval needs a name. An SD bar and a confidence interval answer different questions.

Optional reference: counts to a Python chart

Python example. Every number comes from the small illustrative registration table.

```
import matplotlib.pyplot as plt
passed = [60, 18]
total = [100, 20]
rates = [100 * a / b for a, b in zip(passed, total)]
# rates = [60.0, 90.0]
fig, ax = plt.subplots()
ax.bar(['Full-time (60/100)', 'Part-time (18/20)'], rates)
ax.set_ylim(0, 100)
ax.set_ylabel('Observed pass rate (%)')
ax.set_title('Illustrative pass rates by study mode')
```

First bar: $100 \times 60/100 = 60$. Second bar: $100 \times 18/20 = 90$.

Optional reference: logarithmic position

Optional extension. Base10 logs ask: what power of10 gives this value?

Value x	1	10	100	1,000
Power of10	10^0	10^1	10^2	10^3
Position $\log_{10}(x)$	0	1	2	3

10 to100: ratio $100/10=10$, log distance $2-1=1$
100 to1,000: ratio $1,000/100=10$, log distance $3-2=1$

Equal log distances represent equal multiplication factors. Zero has no finite log position.

References and where the examples come from

Public teaching sources inform the concepts. Numerical examples are labelled as illustrative.

Source	Concept and corresponding example
Stanford CS448B	Encoding and perception: bar baselines, pie angles, bubble area
Harvard CS171	Marks, channels and chart-design critique
Berkeley Data100	Histograms, density and comparisons across groups
UW CSE512	Critique of different claims from the same data

Synthetic registration counts come from Week5. Other examples are constructed for this lecture.