Lecture 3

Visualizing Variation in Numerical Data

§2.1 & §2.2

1. Visualizing Variation in Numerical Data

Distribution of Numerical Data

Definition (Distribution).
The **distribution** of a sample describes the observed values and how frequently they occur.

For example,

$$2, \quad 3, \quad 3, \quad 4, \quad 4, \quad 4, \quad 5, \quad 6, \quad 6, \quad 8.$$

Study Hours	Frequency
2	1
3	2
4	3
5	1
6	2
8	1
Total	**10**

$$\boxed{\text{Distribution} = \text{Values} + \text{Frequencies}}$$

Example: Reading a Distribution

Consider the data representing the number of text messages sent by 10 students during one evening:

$$2, \quad 3, \quad 3, \quad 4, \quad 4, \quad 4, \quad 5, \quad 6, \quad 6, \quad 8.$$

We can organize the data into a frequency table:

Number of Messages	Frequency
2	1
3	2
4	3
5	1
6	2
8	1
Total	**10**

Now we can answer several questions.

a) Which value occurred most often?
The value 4 occurred three times.

$$\boxed{\text{Most frequent value} = 4}$$

b) How many students sent at least 5 messages?

$$1 + 2 + 1 = 4.$$

4 students.

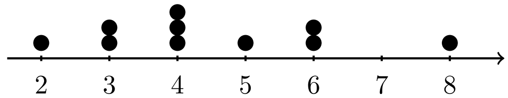
c) How many students sent fewer than 5 messages?

$$1 + 2 + 3 = 6.$$

6 students.

1.1 Dotplots

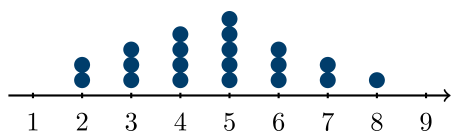
A **dotplot** displays each observation with a dot above its value.



$$\text{Frequency of } 4 = 3.$$

Example: Interpreting a Dotplot

Suppose the dotplot represents the number of hours that 15 students studied before an exam.



Describe the distribution.

- The observations range from about 2 to 8 hours.
- The center appears to be around 4–5 hours.
- The distribution has one main mound.
- There are no obvious gaps.

2. Histograms

A **histogram** displays a numerical distribution by grouping observations into intervals, called **bins** or **classes**. The height of each bar represents the frequency in that interval.

> **Definition**
> A **histogram** groups numerical observations into intervals and uses the height of each bar to show the frequency in each interval.

2.1 Bins and Boundaries

A **bin** is an interval of numerical values. For example,

$$[50, 60), \qquad [60, 70), \qquad [70, 80).$$

The notation

$$[a, b)$$

means

$$a \leq x < b.$$

Thus,

$$60 \in [60, 70) \qquad \text{and} \qquad 70 \in [70, 80).$$

Rule: Bins should not overlap. Every observation must belong to exactly one bin.

Example: Constructing a Histogram from Data

Suppose the following are the waiting times, in minutes, for 20 customers:

$$52,\ 61,\ 64,\ 68,\ 71,\ 72,\ 74,\ 75,\ 76,\ 78,$$
$$81,\ 83,\ 84,\ 85,\ 87,\ 91,\ 93,\ 94,\ 96,\ 98.$$

Three-step method for constructing a histogram

1. Choose intervals of equal width.
2. Count the observations in each interval.
3. Draw touching bars with heights equal to the frequencies.

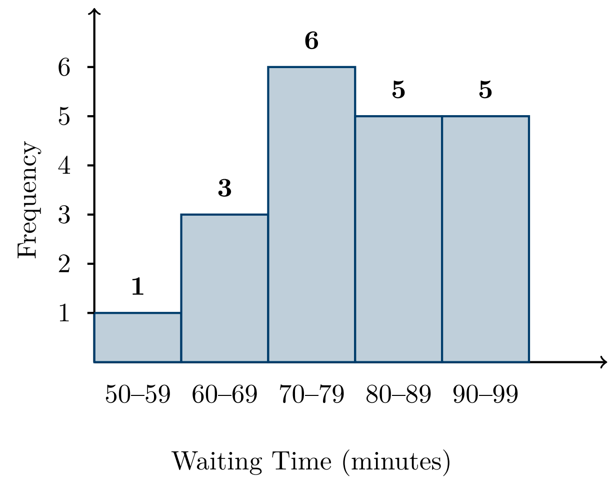
We will group the observations into intervals of width 10:

50–59, 60–69, 70–79, 80–89, 90–99.

Count how many observations fall into each interval.

Waiting Time	Frequency	Example Values
50–59	1	52
60–69	3	61, 64, 68
70–79	6	71, 72, 74, 75, 76, 78
80–89	5	81, 83, 84, 85, 87
90–99	5	91, 93, 94, 96, 98
Total	**20**	

The histogram is obtained by plotting these frequencies.



The height of each bar represents the frequency of observations in that interval. For example, the bar for 70–79 has height 6.

Therefore,

$$\boxed{\text{6 customers waited between 70 and 79 minutes.}}$$

2.2 Histogram vs. Bar Chart

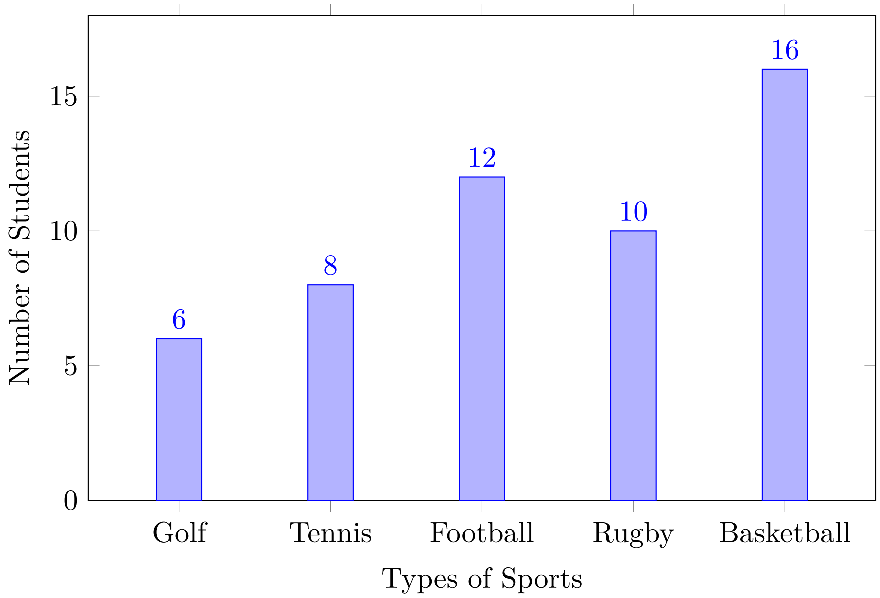
Histogram	**Bar Chart**
Numerical data	Categorical data
Bars touch	Bars are separated
Intervals	Categories

Bar Chart

Question: Let us assume that Rob has taken a survey of his classmates to find which kind of sports they prefer and noted the result in the form of a table.

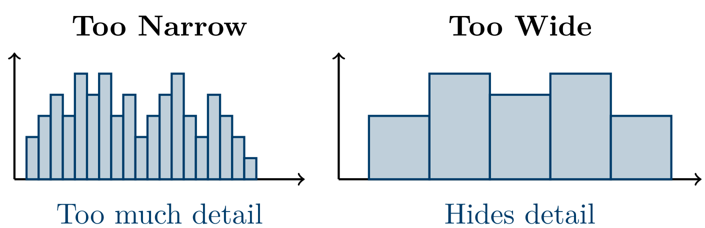
Type of Sport	**Number of Students**
Golf	6
Tennis	8
Football	12
Rugby	10
Basketball	16

Represent the above data using a bar graph.



Bin Width

The bin width affects how much detail a histogram shows.



Good bin width = enough detail without unnecessary noise

2.3 Frequency and Relative Frequency

Suppose we ask 20 students how many books they read during the summer. We group their answers into five intervals.

Books Read	Frequency	Relative Frequency
0–1	2	$\frac{2}{20} = 0.10 = 10\%$
2–3	4	$\frac{4}{20} = 0.20 = 20\%$
4–5	6	$\frac{6}{20} = 0.30 = 30\%$
6–7	4	$\frac{4}{20} = 0.20 = 20\%$
8–9	2	$\frac{2}{20} = 0.10 = 10\%$
Total	**20**	**100%**

A **frequency histogram** uses counts on the vertical axis. A **relative-frequency histogram** uses proportions or percentages on the vertical axis.

For example, 6 students read between 4 and 5 books. Therefore,

$$\frac{6}{20} = 0.30 = 30\%.$$

Thus, the frequency is 6, while the relative frequency is 0.30 or 30%.

Example: Interpreting Relative Frequency

Suppose a survey asks 50 students how many hours they exercise each week.

Hours	Frequency	Relative Frequency
0–2	8	0.16
3–5	17	0.34
6–8	15	0.30
9–11	7	0.14
12–14	3	0.06

Consider the interval 3–5 hours.
There are 17 students in this group, so

$$\frac{17}{50} = 0.34 = 34\%.$$

Therefore:

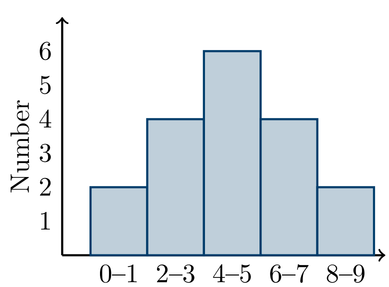
Interpretation:
Approximately **34% of the students exercise between 3 and 5 hours per week**.

Notice that a relative frequency is especially useful when comparing groups of different sizes.
For example, 17 students out of 50 and 34 students out of 100 represent the same proportion:

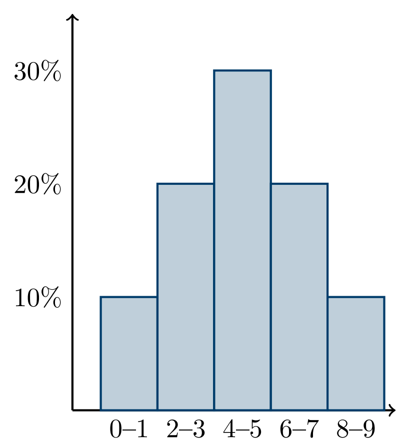
$$\frac{17}{50} = \frac{34}{100} = 0.34.$$

Both groups have a relative frequency of 34%.

Frequency Histogram



Relative-Frequency Histogram



Notice that the **shape of the two histograms is the same**. Only the units on the vertical axis change.

$$\boxed{\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Total Number of Observations}}}$$

3. Stemplots

A **stemplot** displays numerical data by separating each observation into a **stem** and a **leaf**.

> **Definition (Stemplot).**
> A **stemplot** divides each observation into a **stem** and a **leaf**.
> The **leaf** is the last digit, and the **stem** contains the digits preceding it.

Example 1

Consider the data:

$$21, \quad 23, \quad 24, \quad 27, \quad 31, \quad 34, \quad 35, \quad 42, \quad 44.$$

For example,

$$27 = \boxed{2} \mid \boxed{7}, \qquad 34 = \boxed{3} \mid \boxed{4}.$$

Thus, the stem contains the tens digit and the leaf contains the ones digit.

Stem	**Leaves**
2	1 3 4 7
3	1 4 5
4	2 4

$$\boxed{\text{Key: } 2 \mid 7 = 27}$$

Example 2

Suppose the data are

$$2, \ 4, \ 7, \ 11, \ 13, \ 15, \ 18, \ 21, \ 24, \ 27.$$

A stemplot is:

Stem	Leaves
0	2 4 7
1	1 3 5 8
2	1 4 7

The key is

$$\boxed{1|3 = 13}$$

Here, the **stem** gives the tens digit and the **leaf** gives the ones digit. The stem 0 represents values from 0 through 9.

Building a Stemplot Step by Step

Consider the following quiz scores:

$$12, \ 15, \ 17, \ 18, \ 21, \ 22, \ 24, \ 27,$$
$$31, \ 32, \ 34, \ 36, \ 38, \ 41, \ 44.$$

Step 1: Identify the stem and leaf.
For the observation 27,

$$27 = 2 \mid 7.$$

The stem is 2, and the leaf is 7.
Step 2: List the stems.
The scores range from the 10's through the 40's, so we use

$$1, \quad 2, \quad 3, \quad 4.$$

Step 3: Place each leaf next to its stem.

Stem	**Leaves**				
1	2	5	7	8	
2	1	2	4	7	
3	1	2	4	6	8
4	1	4			

$$\boxed{\text{Key: } 2 \mid 7 = 27}$$

Step 4: Check the number of leaves.
There are

$$4 + 4 + 5 + 2 = 15$$

leaves, matching the 15 original observations.

Useful check:
The total number of leaves in a stemplot should equal the total number of observations in the original data set.

Reading a Stemplot

Each leaf represents one observation.
For example,

$$3 \mid 1, \quad 3 \mid 4, \quad 3 \mid 5$$

represent

$$31, \quad 34, \quad 35.$$

3.1 Ordering the Leaves

Leaves should be arranged from smallest to largest:

$$\begin{array}{r|llll} 2 & 1 & 3 & 4 & 7 \\ 3 & 1 & 4 & 5 & \\ 4 & 2 & 4 & & \end{array}$$

Example 4

A professor records the exam scores of 40 students. The scores are:

$$\begin{aligned}&42,\ 45,\ 47,\ 48,\ 51,\ 52,\ 54,\ 55,\ 56,\ 58,\\&61,\ 62,\ 63,\ 65,\ 66,\ 67,\ 68,\ 69,\ 71,\ 72,\\&73,\ 74,\ 75,\ 76,\ 78,\ 79,\ 81,\ 82,\ 83,\ 84,\\&85,\ 86,\ 87,\ 88,\ 89,\ 91,\ 92,\ 94,\ 96,\ 98.\end{aligned}$$

Construct a stemplot.

Stem	**Leaves**
4	2 5 7 8
5	1 2 4 5 6 8
6	1 2 3 5 6 7 8 9
7	1 2 3 4 5 6 8 9
8	1 2 3 4 5 6 7 8 9
9	1 2 4 6 8

Questions

a) How many students scored between 50% and 69%, inclusive?
Count the leaves in stems 5 and 6:

$$6 + 8 = 14.$$

14 **students.**

b) How many students scored at least 80%?
Count the leaves in stems 8 and 9:

$$9 + 5 = 14.$$

14 **students.**

c) What was the highest exam score?
Look at the largest stem and the largest leaf:

$$9 \mid 8 = 98.$$

98**.**

d) What was the lowest exam score?
Look at the smallest stem and the smallest leaf:

$$4 \mid 2 = 42.$$

42**.**

e) How many students scored in the 70's?
Count the leaves in stem 7:

$$1\ 2\ 3\ 4\ 5\ 6\ 8\ 9$$

There are 8 leaves.
8 **students.**