# Frequency Distributions

A group of 17 students were asked to choose their favourite course from among Art, Literature, or Science. Their responses are listed below:

| | | | | | |
|---|---|---|---|---|---|
| A | S | A | S | L | S |
| S | L | L | A | A | A |
| A | A | L | A | S | |

**Q:** What conclusions can you draw from this data?

**A:** Not many! We need some way to organize the data.

# Frequency Tables

A **frequency table** is a way to organize data which can be split into **categories** by counting the number data points in each category (frequency).

We had the following data:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | S | A | S | L | S | S | L | L |
| A | A | A | A | A | L | A | S | |

We can split this into 3 categories: Art, Literature, and Science, and we count the number of responses for each.

Art appears 8 times, Literature 4, and Science 5.

| Class preference | # Students |
|---:|---|
| Art | 8 |
| Literature | 4 |
| Science | 5 |

# Frequency Tables

**Example:** A small private university is renowned for its Science and Engineering program. The following frequency table summarizes the students' major by subject:

| Major | # Students |
|---:|---|
| Art | 94 |
| Biology | 303 |
| Chemistry | 334 |
| Engineering | 412 |
| Languages | 327 |
| Math | 238 |
| Physics | 348 |
| Social Sciences | 377 |
| Total | 2433 |

# Numerical Data

Sometimes, the data we are dealing with is numerical (exam scores, weights, income, etc...).

**Example:** The following data represents the final exam scores from a calculus class of 29 students:

| | | | | | |
|---|---|---|---|---|---|
| 81 | 76 | 85 | 72 | 94 | 67 |
| 81 | 90 | 84 | 92 | 75 | 66 |
| 79 | 69 | 68 | 72 | 68 | 80 |
| 84 | 55 | 79 | 75 | 73 | 76 |
| 92 | 100 | 95 | 66 | 95 | |

**Q:** How do we build a frequency table for this type of data?

**A:** We create categories by constructing **bins** or intervals, and count the number of data points in each interval.

# Numerical Data

**Example:** The following data represents the final exam scores from a calculus class of 29 students:

| | | | | | |
|---|---|---|---|---|---|
| 81 | 76 | 85 | 72 | 94 | 67 |
| 81 | 90 | 84 | 92 | 75 | 66 |
| 79 | 69 | 68 | 72 | 68 | 80 |
| 84 | 55 | 79 | 75 | 73 | 76 |
| 92 | 100 | 95 | 66 | 95 | |

We can create the bins by letter grade (0-59 is F, 60-69 is D, etc...). This gives the following frequency table:

| Letter Grade | # Students |
|---|---|
| A | 7 |
| B | 6 |
| C | 9 |
| D | 6 |
| F | 1 |
| Total | 29 |

# Numerical Data

If the nature of the data doesn't make it clear how to determine the bins, we can split the entire data range into **intervals of equal length**:

- Locate the largest and smallest data point
- Divide the difference between them by the number of desired intervals to get the **interval length**
- Adjust this value to a relatively simple number

**Note:** sometimes it may be helpful to pick a value smaller/larger than the above, just so the numbers come out nicer.

# Numerical Data

**Example:** The following represent amounts spent (in USD) by 14 students on supplies for an art class:

| | | | | | | |
|---|---|---|---|---|---|---|
| 69 | 81 | 83 | 41 | 103 | 94 | 53 |
| 96 | 135 | 56 | 81 | 58 | 56 | 52 |

The smallest value is 41, and the largest is 135. So the numbers are nicer, we choose to go between 40 (slightly below 41) and 140 (slightly above 135), and we will use 5 intervals.

To get the interval length, $140 - 40 = 100$, and $100/5 = 20$, so each interval will be of length 20. The bins are going to be:

$40 - 60 \qquad 60 - 80 \qquad 80 - 100 \qquad 100 - 120 \qquad 120 - 140$

# Numerical Data

Data:

| | | | | | | |
|---|---|---|---|---|---|---|
| 69 | 81 | 83 | 41 | 103 | 94 | 53 |
| 96 | 135 | 56 | 81 | 58 | 56 | 52 |

Intervals:

$40 - 60 \quad 60 - 80 \quad 80 - 100 \quad 100 - 120 \quad 120 - 140$

We obtain the following frequency table:

| Price Range | Freq |
|---|---|
| \$40 — \$60 | 6 |
| \$60 — \$80 | 1 |
| \$80 — \$100 | 5 |
| \$100 — \$120 | 1 |
| \$120 — \$140 | 1 |
| Total | 14 |

# Relative Frequencies

Instead of specifying the frequency (count) of data points in each bin, we can choose to specify the **relative frequency**. This is obtained by taking the frequency and dividing by the total number of data points.

The advantage of using relative frequencies: they provide an **estimate of the probability** that a randomly chosen data point will fall in the corresponding category/bin.

**Example:** Class preferences of students:

| Class preference | Freq | Rel Freq |
|---:|:---:|:---:|
| Art | 8 | 0.47 |
| Literature | 4 | 0.24 |
| Science | 5 | 0.29 |
| Total | 17 | 1 |

# Relative Frequencies

**Example:** Student majors by subject.

| Major | Freq | Rel Freq |
|---|---|---|
| Art | 94 | 0.038 |
| Biology | 303 | 0.125 |
| Chemistry | 334 | 0.137 |
| Engineering | 412 | 0.170 |
| Languages | 327 | 0.134 |
| Math | 238 | 0.098 |
| Physics | 348 | 0.143 |
| Social Sciences | 377 | 0.155 |
| Total | 2433 | 1 |

# Histograms

While frequency tables are useful, sometimes it is more convenient to have a **visual representation** of the data. A **histogram** enables us to translate a frequency table into a graph.

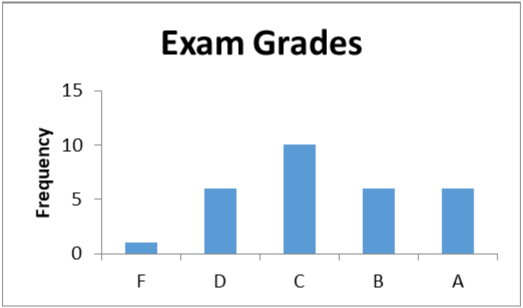
The $x$-**axis** will be labeled using the categories, and the $y$-**axis** represents the frequency (or relative frequency) of the corresponding category.

# Histograms

**Example:** Student grades on an exam

| | | | | | |
|---|---|---|---|---|---|
| 81 | 76 | 85 | 72 | 94 | 67 |
| 81 | 90 | 84 | 92 | 75 | 66 |
| 79 | 69 | 68 | 72 | 68 | 80 |
| 84 | 55 | 79 | 75 | 73 | 76 |
| 92 | 100 | 95 | 66 | 95 | |

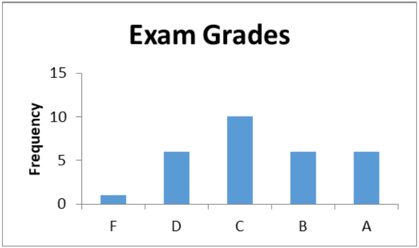
| Grade | F | RF |
|---|---|---|
| A | 7 | 0.24 |
| B | 6 | 0.21 |
| C | 9 | 0.31 |
| D | 6 | 0.21 |
| F | 1 | 0.03 |
| Total | 29 | 1 |

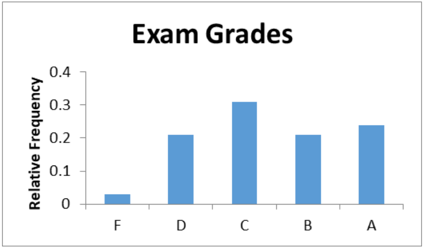


# Histograms

**Q:** What happens if we plot relative frequency instead of frequency?

| Grade | F | RF |
|---|---|---|
| A | 7 | 0.24 |
| B | 6 | 0.21 |
| C | 9 | 0.31 |
| D | 6 | 0.21 |
| F | 1 | 0.03 |
| Total | 29 | 1 |



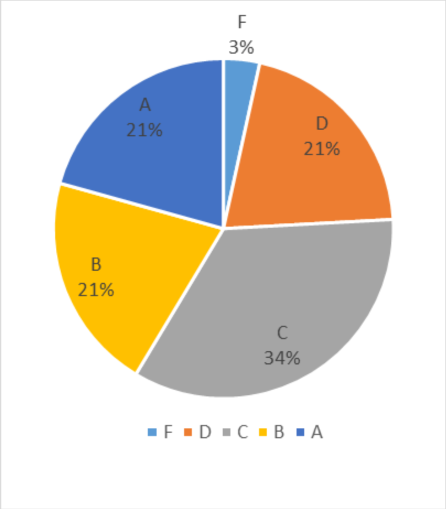

**A:** The histogram will have the **same shape**, but **different scaling** on the $y$-axis.

# Pie Charts

A **pie chart** is a tool which helps us better visualize each category and what **proportion of the whole** they occupy.

**Example:** Exam grades.

| Grade | F | RF |
|---|---|---|
| A | 7 | 0.24 |
| B | 6 | 0.21 |
| C | 9 | 0.31 |
| D | 6 | 0.21 |
| F | 1 | 0.03 |
| Total | 29 | 1 |



# Pie Charts

The categories in a pie chart may be **color coded**, and each color explained in the **legend**. Alternately, we may directly label each chunk with the categories directly on the chart.