Today's plan:

- We start Chapter 4. Section 4.1.1: Data sets, and Section 4.1.2: Frequency tables, bar graphs, piecharts, histograms.

Definition

Statistics is the science of the collection, organization, and interpretation of data.

We begin with:

We begin with:

- different graphical ways to present data

We begin with:

- different graphical ways to present data
- numerical summaries of data

Section 4.1.1: Data sets.

Example

Database for the DMV (Department of Motor Vehicles)

Example

Database for the DMV (Department of Motor Vehicles)

- There's a DMV record for each registered driver.

Example

- Each DMV record has a number of fields like:

id number	date of birth
last name	first name
address	gender
eye color	height
issue date	expiration date

Example

- The structure of all the records is the same, but the information in them varies from one record to another.

We'll use DMV records to show some basic concepts about data sets.

We'll use DMV records to show some basic concepts about data sets.

Definition

A **data set** is a collection of **data points** that carry information about a certain **population**.

In the DMV example

- each record is a data point

In the DMV example

- each record is a data point
- the whole DMV database is a data set

In the DMV example

- each record is a data point
- the whole DMV database is a data set
- the population consists of all registered drivers

The data in each data point may be

- quantitative (or numerical) data

The data in each data point may be

- quantitative (or numerical) data
- qualitative (or non-numerical) data

The data in each data point may be

- quantitative (or numerical) data
- qualitative (or non-numerical) data
- combination of both

The data in each data point may be

- quantitative (or numerical) data
- qualitative (or non-numerical) data
- combination of both

Definition

Each piece of a data point is called a coordinate.

For DMV records, each data point has a combination of quantitative and qualitative data.

For DMV records, each data point has a combination of quantitative and qualitative data.

- *height* is a numerical coordinate

For DMV records, each data point has a combination of quantitative and qualitative data.

- *height* is a numerical coordinate
- *last name* is a non-numerical coordinate

Coordinates of a data point are examples of variables.

Coordinates of a data point are examples of variables.

Definition

A variable is any quantity or characteristic whose value can change.

Coordinates of a data point are examples of variables.

Definition

A variable is any quantity or characteristic whose value can change.

New variables arise by adding, subtracting, or combining other ones (e.g., "height plus weight").

Definition

The size of a data set is the number of data points in it.

Definition

The size of a data set is the number of data points in it.

Usually we denote the size of a data set by either **N** or **n**.

Example

- The NY state DMV data set has a size in the millions.

Example

- The NY state DMV data set has a size in the millions.
- The Library of Congress has over 21 million cataloged books.

Example

Home Town Pharmacy's sales for the week were:

S	M	T	W	R	F	S
$2,548,	$1,225,	$1,732,	$1,871,	$975,	$2,218,	$1,339.

This data set has size $n = 7$.

This data set has size $n = 7$.
Each data point has

- a numerical coordinate (sales)
- a non-numerical coordinate (day of the week)

Table: Math 130 First Test Scores (previous semester)

ID	Score	ID	Score	ID	Score	ID	Score	ID	Score
0104	78	6607	88	1491	65	8882	62	2026	55
8101	55	9046	91	7959	31	3920	24	6293	62
5036	85	6764	84	6337	77	2193	95	0980	80
2639	83	1769	75	7935	84	1364	50	9130	62
7380	40	8138	69	2737	73	0467	56	4623	80
9163	93	4257	76	6187	81	4516	78	1269	23
0816	64	3891	61	7435	89	9884	34	9842	55
4022	82	5267	73	9620	71	2860	99	2828	63
9802	95	9393	64	6461	84	7439	78	1878	83
6835	58	0550	94	5178	79	7348	88	8416	90
9374	73	9027	92	6145	67	2640	86	4261	99
9728	75	2007	82	4031	57	3590	90	0620	61
3621	73	3885	68	9533	70	2306	20	7387	98
6228	87	9554	82	3467	74	6990	88	2778	95
2458	94	9098	68	6794	66	2508	71	5081	68
2627	91	9449	91	3302	82	8054	99	7077	87
0341	61	1176	83	0573	96	4671	93	2883	81
7538	61	1085	67	0387	57	3962	77	6869	80
3902	58	8708	86	8920	71	1530	89	2708	64
1235	96	8870	39	4790	60	3068	74	7277	67
3403	58	6448	81	1924	89	7033	92	6956	60
4551	64	8750	65	5381	92	0508	54	8538	65
4997	81	4490	77	8626	93	2090	72	7227	86
1285	95	9883	97	8795	79	8362	91	4920	85
7897	55	2878	63	3427	76	9766	82	0072	86
0841	73	3628	77	3997	74	0749	81	4303	73
2365	56	6229	90	5027	76	5710	85	3353	74

This data set has size $N = 135$.

This data set has size $N = 135$.
Each data point has two coordinates:

- **ID** number (numeric)
- the **score** (numeric)

This data set has size $N = 135$.
Each data point has two coordinates:

- **ID** number (numeric)
- the **score** (numeric)

For the **ID**, we could have also used a non-numeric data field.
In other words, that field is not intrinsically numeric.

On the other hand, the **score**, is intrinsically numeric: a larger score means a higher grade.

The data points here are listed randomly. That's typical for raw data.

The data points here are listed randomly. That's typical for raw data.

- The lack of order makes it hard to see how people generally did on the test.

The data points here are listed randomly. That's typical for raw data.

- The lack of order makes it hard to see how people generally did on the test.
- One of the goals of **descriptive statistics** is to organize the data set.

Ways to organize a data set:

- sort the data set

Ways to organize a data set:

- sort the data set
- compile a frequency table

Ways to organize a data set:

- sort the data set
- compile a frequency table
- draw a bar graph

Ways to organize a data set:

- sort the data set
- compile a frequency table
- draw a bar graph
- draw a piechart

Ways to organize a data set:

- sort the data set
- compile a frequency table
- draw a bar graph
- draw a piechart
- draw a histogram

Section 4.1.2: Frequency tables, bar graphs, piecharts, and histograms.

Sorting and Frequency Tables

Definition

A sorted data set is the result of sorting the data set in ascending or descending order.

In a class with 95 students, a 25 point quiz is given.

ID	Score	ID	Score	ID	Score	ID	Score	ID	Score
6529	4	5128	12	9637	14	2718	15	5774	18
5204	5	5626	12	5962	14	7262	15	7514	18
6265	8	6948	12	7028	14	8592	16	9767	18
6283	8	8062	12	2894	14	8254	16	4901	18
8771	9	2215	13	6602	14	5075	16	5914	18
5541	9	3063	13	1635	14	5754	16	9050	18
2291	10	4048	13	6360	14	1319	16	8156	18
8847	10	3020	13	7688	14	9741	16	8941	19
4290	10	6594	13	2684	15	2885	16	8985	19
3685	11	9184	13	2774	15	3372	16	6825	19
3257	11	1461	13	2048	15	4179	16	4183	19
1640	11	7162	13	1568	15	4557	17	4632	19
7254	11	3065	13	9283	15	6715	17	5895	20
7114	11	5666	13	4255	15	1168	17	7797	20
2086	12	9779	13	4026	15	6871	17	3224	20
2049	12	5700	13	3583	15	1096	17	4149	21
9062	12	4437	14	8669	15	3062	17	2872	21
7008	12	5320	14	5242	15	3473	17	5331	22
1246	12	4343	14	5646	15	1731	17	1821	25

The quiz scores are sorted in ascending order, based on the second coordinate.

The quiz scores are sorted in ascending order, based on the second coordinate.

Sorting is the first step in organizing the data set. Once we have it sorted, we can compile the frequency table

score	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25
freq.	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1
cum fr.	1	2	4	6	9	14	23	35	46	59	68	76	83	88	91	93	94	95

score	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25
freq.	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1
cum fr.	1	2	4	6	9	14	23	35	46	59	68	76	83	88	91	93	94	95

- The first row contains the scores.

score	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25
freq.	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1
cum fr.	1	2	4	6	9	14	23	35	46	59	68	76	83	88	91	93	94	95

- The first row contains the scores.
- The frequency row says how many quizzes have that score.

score	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25
freq.	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1
cum fr.	1	2	4	6	9	14	23	35	46	59	68	76	83	88	91	93	94	95

- The first row contains the scores.
- The frequency row says how many quizzes have that score.
- The cumulative frequency row keeps a running total of the frequencies.

score	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25
freq.	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1
cum fr.	1	2	4	6	9	14	23	35	46	59	68	76	83	88	91	93	94	95

- The first row contains the scores.
- The frequency row says how many quizzes have that score.
- The cumulative frequency row keeps a running total of the frequencies.
- Note: The last entry of the third row gives the size of the data set.

Frequency tables are convenient for large data sets with many repeated values.

Bar Graphs

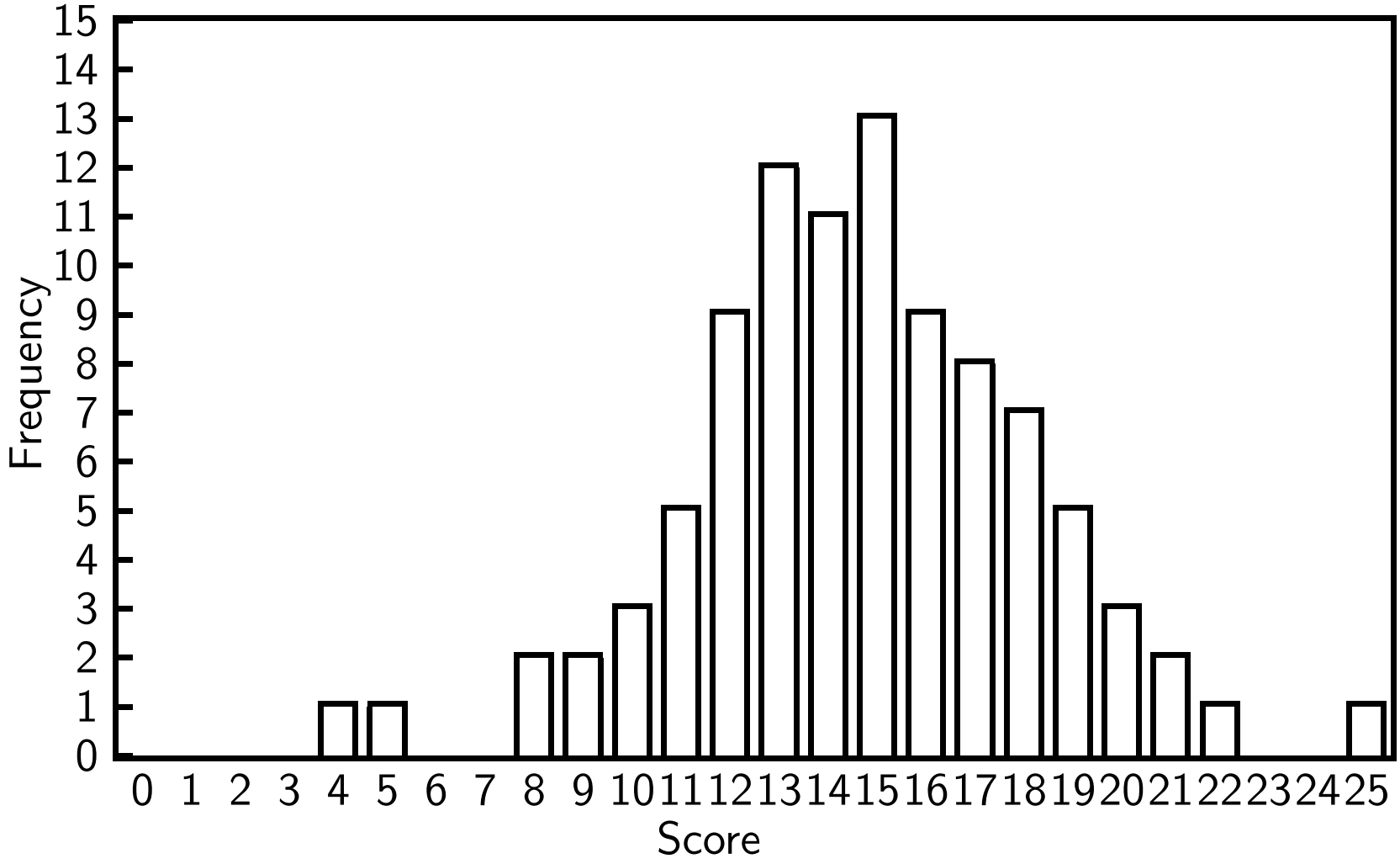
With the frequency table done, we can draw a bar graph.

With the frequency table done, we can draw a bar graph.

- The values of the *score row* are plotted horizontally.

With the frequency table done, we can draw a bar graph.

- The values of the *score row* are plotted horizontally.
- Above each score we draw a rectangle whose height is the corresponding frequency.


Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Score
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Definition

When the number of possible values of a variable is very large, it is convenient to group them.

Definition

When the number of possible values of a variable is very large, it is convenient to group them.
For numerical variables, they are grouped in intervals, which are called class intervals.

Notation

$[a, b)$ will represent the interval of numbers from a to b, including a (square bracket), and excluding b (round parenthesis).

Notation

$[a, b)$ will represent the interval of numbers from a to b, including a (square bracket), and excluding b (round parenthesis).
Similarly,

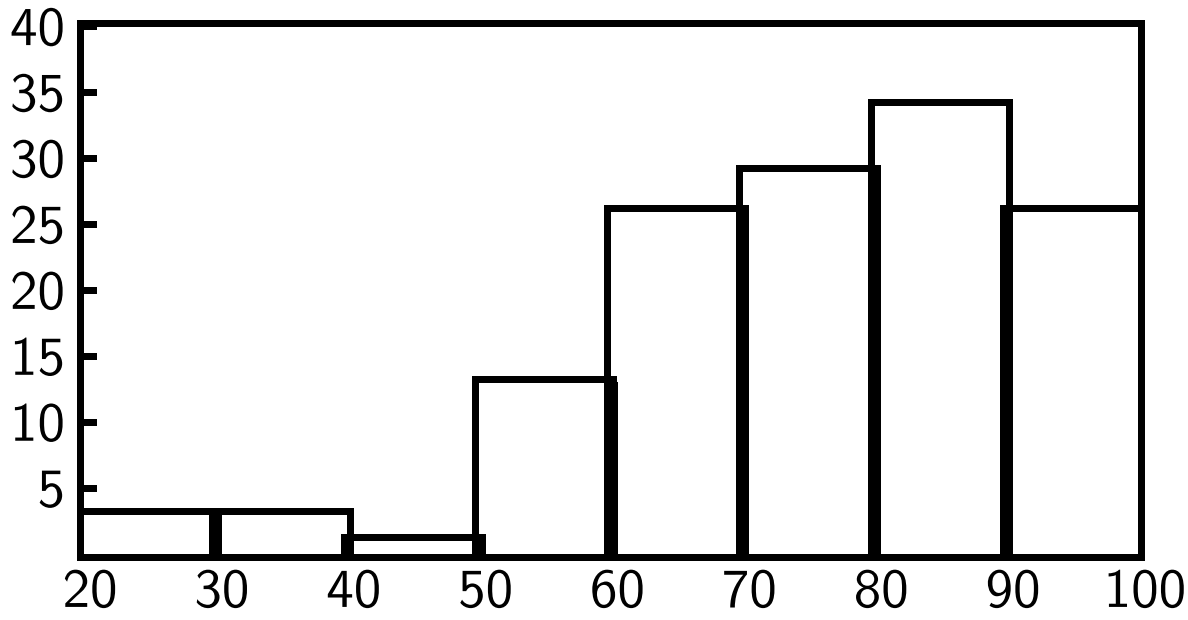
- *$[a, b]$ includes both a and b*
- *(a, b) excludes both a and b*

A bar graph where class intervals aren't used can be drawn either with its bars touching or not, but a bar graph using class intervals should be drawn with the bars touching.

Example

Recall the Math 130 test takes values ranging from 20 to 99.
That's too many values for a bar graph. Group them in intervals of length 10, and then draw the bar graph.

Score	Freq.
$[20, 30)$	3
$[30, 40)$	3
$[40, 50)$	1
$[50, 60)$	13
$[60, 70)$	26
$[70, 80)$	29
$[80, 90)$	34
$[90, 100]$	26



Pie Charts

Example

Database of enrollment in the six schools of a college. The frequency table is....

Example

College	Enrollment	
Science	1,450	22.31%
Humanities	1,980	30.46%
Engineering	950	14.62%
Nursing	320	4.92%
Business	1,250	19.23%
Art	550	8.46%

In addition to the usual columns for data value and frequency, we have a third column for the relative frequency.

In addition to the usual columns for data value and frequency, we have a third column for the relative frequency.

Definition

The relative frequency is the frequency as a percentage of the total.

There are 1,450 students in the College of Science, out of a total of 6,500 students, so the relative frequency for that college is

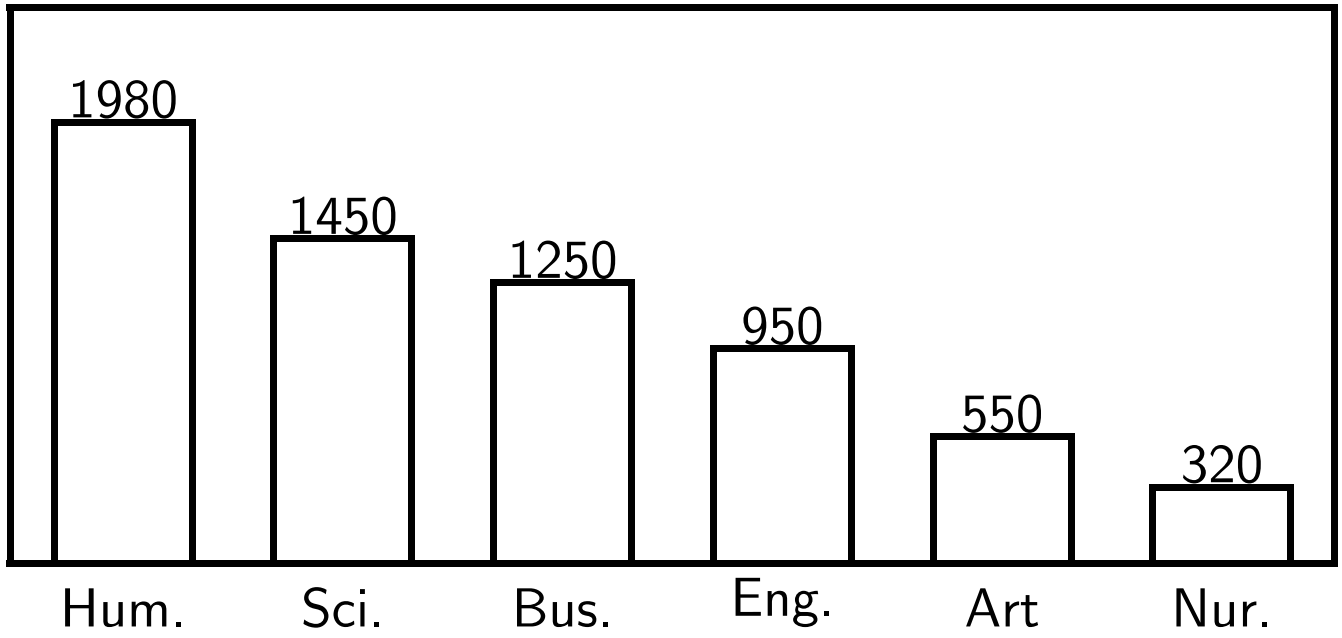
$$\frac{1,450}{6,500} \approx 22.31\%$$

There are 1,450 students in the College of Science, out of a total of 6,500 students, so the relative frequency for that college is

$$\frac{1,450}{6,500} \approx 22.31\%$$

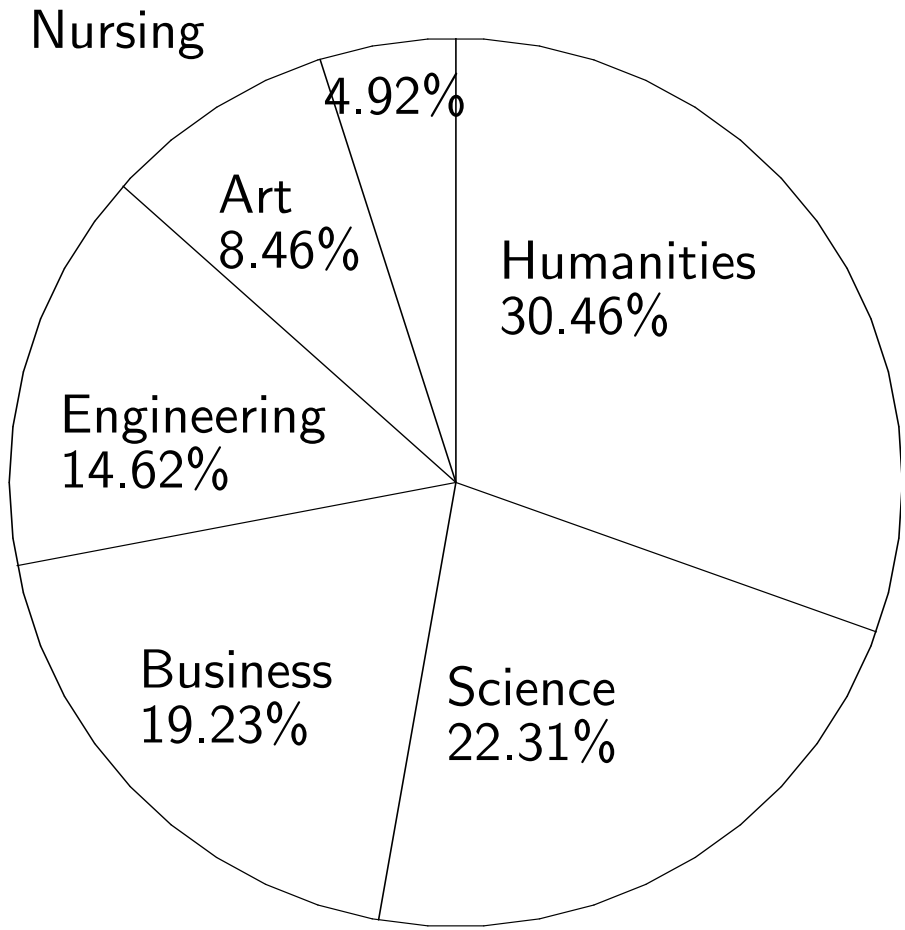
Similar calculations give the percentages for the other colleges, which appear in the third column.

In this case the data is intrinsically non-numeric, so when drawing the bar graph, we can arrange the schools on the horizontal axis arbitrarily.



Here we have arranged them in descending order of enrollment, but other criteria could have been used.

For non-numerical data it is sometimes convenient to arrange the data in a piechart:


Nursing
4.92%
Art
8.46%
Humanities
30.46%
Engineering
14.62%
Business
19.23%
Science
22.31%

- Each value in the data set is represented by a slice.

- Each value in the data set is represented by a slice.
- The size of each slice is proportional to the frequency of that value.

For example, if a frequency occurs 25% of the time, then the slice associated has angle

$$\frac{25}{100} \times 360^\circ = 90^\circ$$

Histograms

- A histogram is similar to a bar graph.

- A histogram is similar to a bar graph.
- In a bar graph, the **height** of a bar is proportional to the percentage.

- A histogram is similar to a bar graph.
- In a bar graph, the **height** of a bar is proportional to the percentage.
- In a histogram, the **area** of a bar is proportional to the percentage.

- A histogram is similar to a bar graph.
- In a bar graph, the **height** of a bar is proportional to the percentage.
- In a histogram, the **area** of a bar is proportional to the percentage.
- When class intervals are all the same size, the histogram looks identical to the bar graph.

Example

A survey at the Clearview textile mill showed that 40% of the workers made up to $10,000 a year and the remaining 60% made more than $10,000, but not more than $50,000. Draw the histogram.

Solution

- *We take $1,000 as a unit of the horizontal axis.*

Solution

- *We take \$1,000 as a unit of the horizontal axis.*
- *The rectangle over the base from 0 to 10,000 should have an area of 40, that is* $10 \times a = 40$, *so* $a = 4\%$ *per 1,000.*

Solution

- *We take \$1,000 as a unit of the horizontal axis.*
- *The rectangle over the base from 0 to 10,000 should have an area of 40, that is* $10 \times a = 40$, *so* $a = 4\%$ *per 1,000.*
- *The rectangle over the base from 10,000 to 50,000 should have an area of 60, that is* $40 \times b = 60$, *so* $b = 1.5\%$ *per 1,000.*

Solution

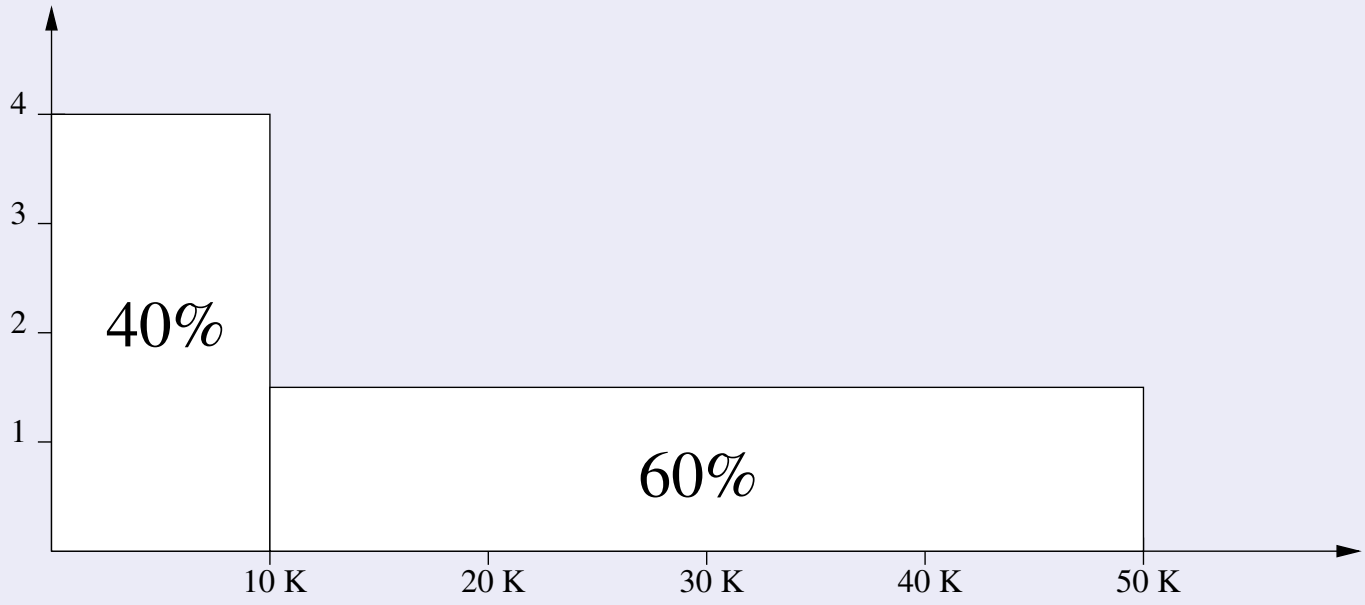


Figure: Histogram for wages at Clearview textile mill

Remarks:

- In a histogram the scale of the vertical axis is "% frequency per unit", whereas in a bar graph it is just the frequency.

Remarks:

- In a histogram the scale of the vertical axis is "% frequency per unit", whereas in a bar graph it is just the frequency.
- a histogram is an estimate, since we do not have any information about the distribution within a block.

- We always assume uniform distribution there.

Example

Estimate the maximum wage of the workers at the Clearview textile mill in the lowest 10% of the pay scale.

Example

Estimate the maximum wage of the workers at the Clearview textile mill in the lowest 10% of the pay scale.

Solution

The lower 40% make up to $10,000.

Example

Estimate the maximum wage of the workers at the Clearview textile mill in the lowest 10% of the pay scale.

Solution

The lower 40% make up to \$10,000. So the lowest 10% make up to $\frac{1}{4} \cdot 10,000 = \$2,500.$

Example

Estimate the minimum wage of the upper 50% on this pay scale.

Solution

- *The lower 50% consists of the 40%, who make up to \$10,000, plus those in the upper 60% who make not more than* $10,000 + \frac{1}{6} \cdot 40,000$.

Solution

- *The lower 50% consists of the 40%, who make up to \$10,000, plus those in the upper 60% who make not more than* $10,000 + \frac{1}{6} \cdot 40,000$.
- *So the upper 50% make at least \$16,666 per year.*

Example

Based on the grade distribution for Test 1 in Math 130 (previous example), draw the histogram for this grade distribution. **Note:** there were 292 students taking Test 1.

grade	class intervals	frequency	%	base	height
A-range	[90,100]	131	44.8	10	4.48
B-range	[78,90)	86	29.4	12	2.4
C-range	[62,78)	46	15.7	16	1.0
D-range	[50,62)	11	3.7	12	0.31
F-range	[0,50)	18	6.1	50	0.12

- The frequency for the interval [90,100] is 131. This means that

$$\frac{131}{292} \times 100 = 44.8\%$$

of all students have grades in the A range.

- The rectangle over the interval [90,100] has an area of 44.8 units and a base of 10.

- The rectangle over the interval [90,100] has an area of 44.8 units and a base of 10. Hence, its height is

$$\frac{44.8}{10} = 4.48$$

- The rectangle over the interval [90,100] has an area of 44.8 units and a base of 10. Hence, its height is

$$\frac{44.8}{10} = 4.48$$

- Similar calculations lead to the heights of the other rectangles.

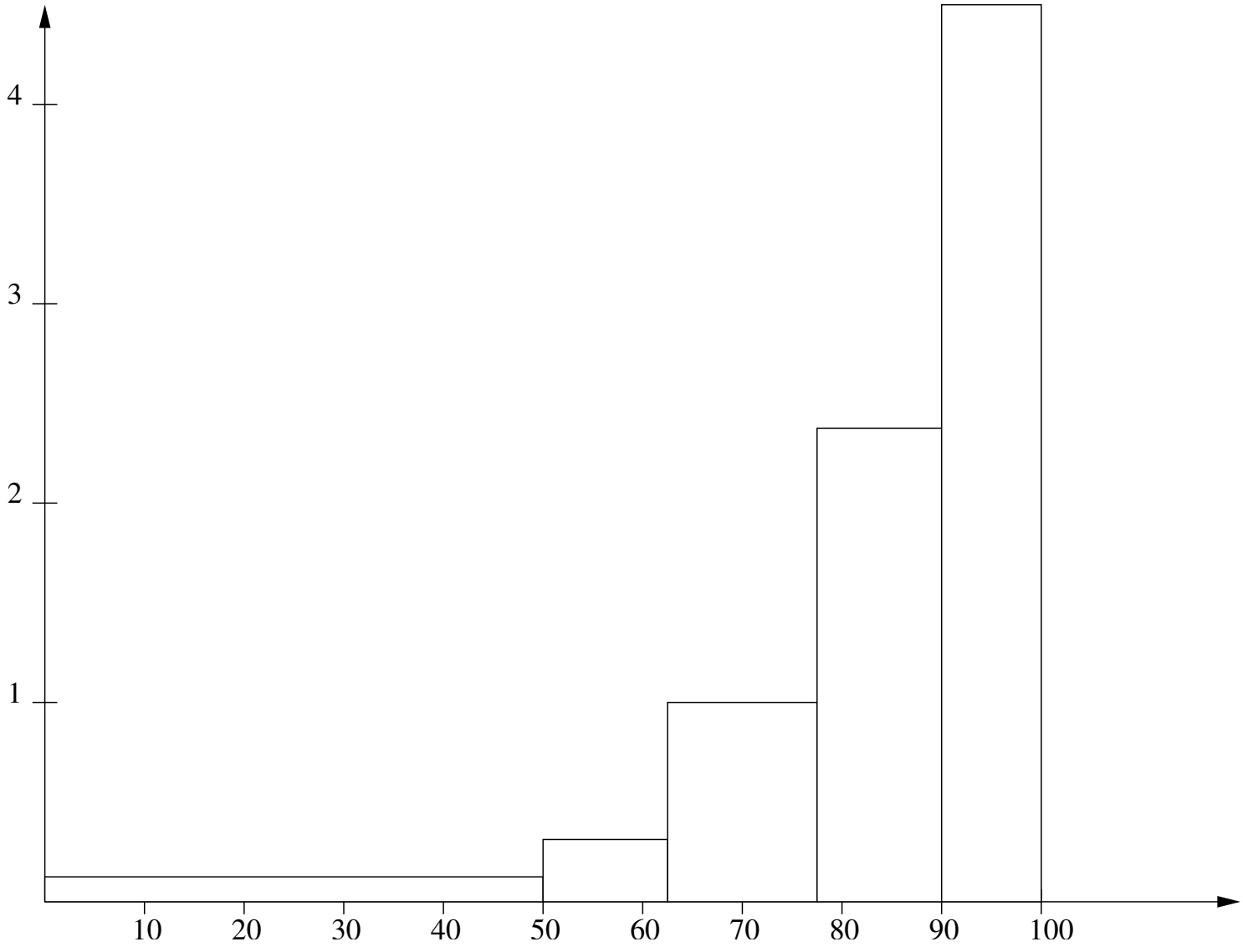
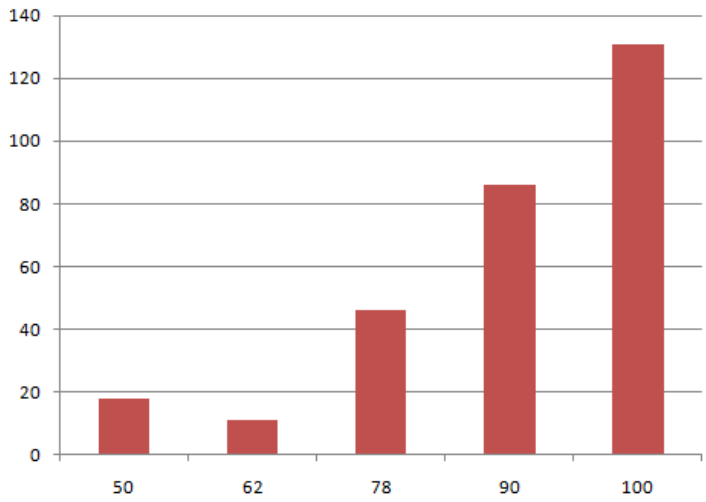
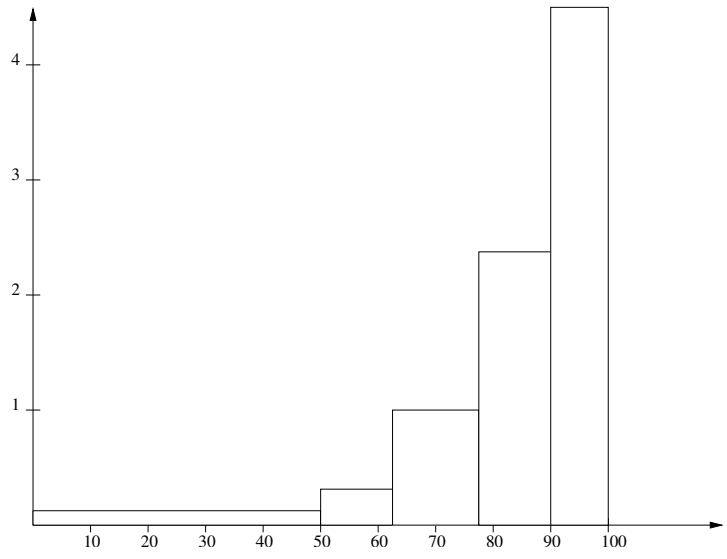


Figure: Histogram for grade distribution for Test I

Let's compare a bar graph with class intervals to a histogram.



Bar Graph



Histogram