ytubiomed@gmail.com

2-Biostatistic

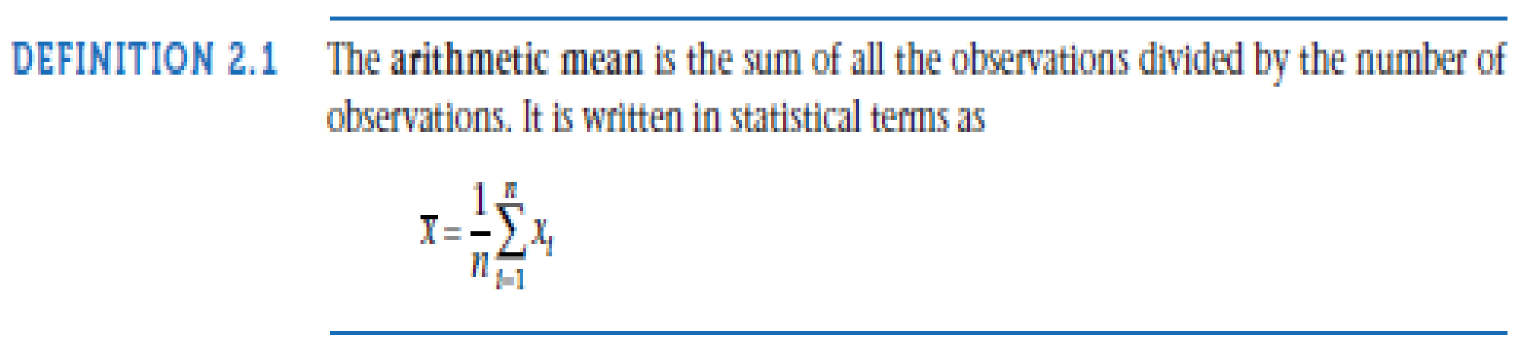
Start: 09:15

Colloquially: halk dilinde

The Arithmetic Mean

- One measure of location for any sample is the arithmetic mean (colloquially called the *average*).

- The arithmetic mean (or mean or sample mean) is usually denoted by *x*.



DEFINITION 2.1 The **arithmetic mean** is the sum of all the observations divided by the number of observations. It is written in statistical terms as

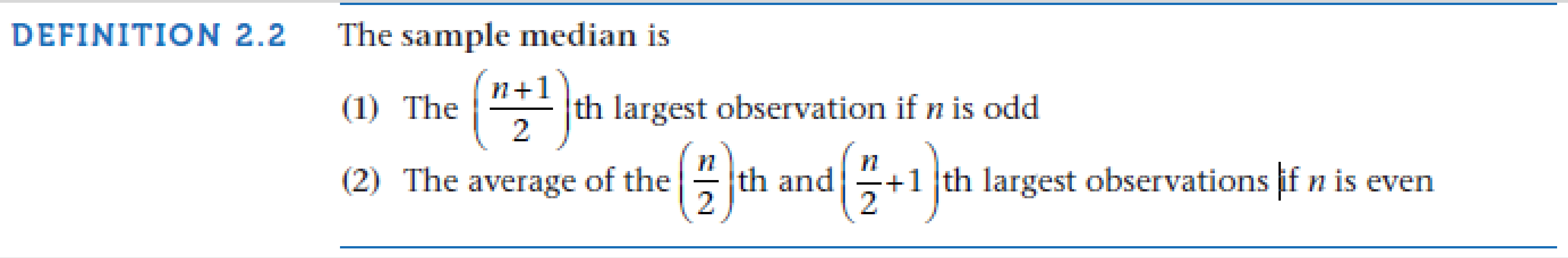
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The Median

order: sıraya koymak
Odd: tek
Even:çift

- An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the median.

- Suppose there are *n* observations in a sample.
- If these observations are ordered from **smallest to largest**, then the median is defined as follows:



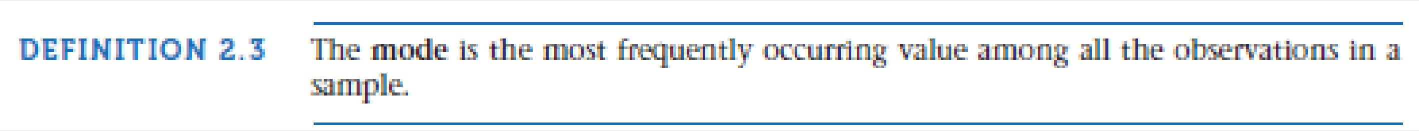
DEFINITION 2.2 The sample median is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The Mode

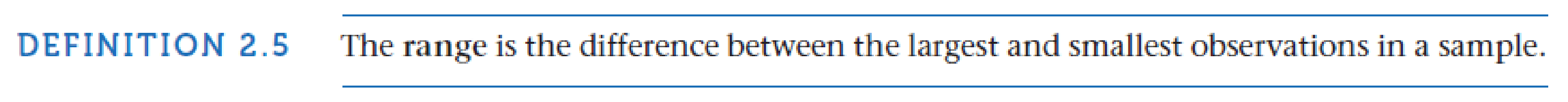
- Another widely used measure of location is the mode.



DEFINITION 2.3 The **mode** is the most frequently occurring value among all the observations in a sample.

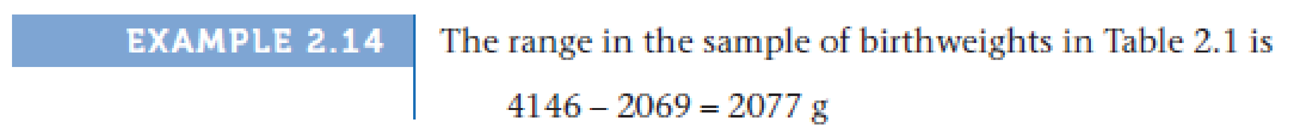
The Range

- Several different measures can be used to describe the variability of a sample.
- Perhaps the simplest measure is the range.



DEFINITION 2.5 The range is the difference between the largest and smallest observations in a sample.

Range=Max-min



EXAMPLE 2.14 The range in the sample of birthweights in Table 2.1 is

$$4146 - 2069 = 2077 \text{ g}$$

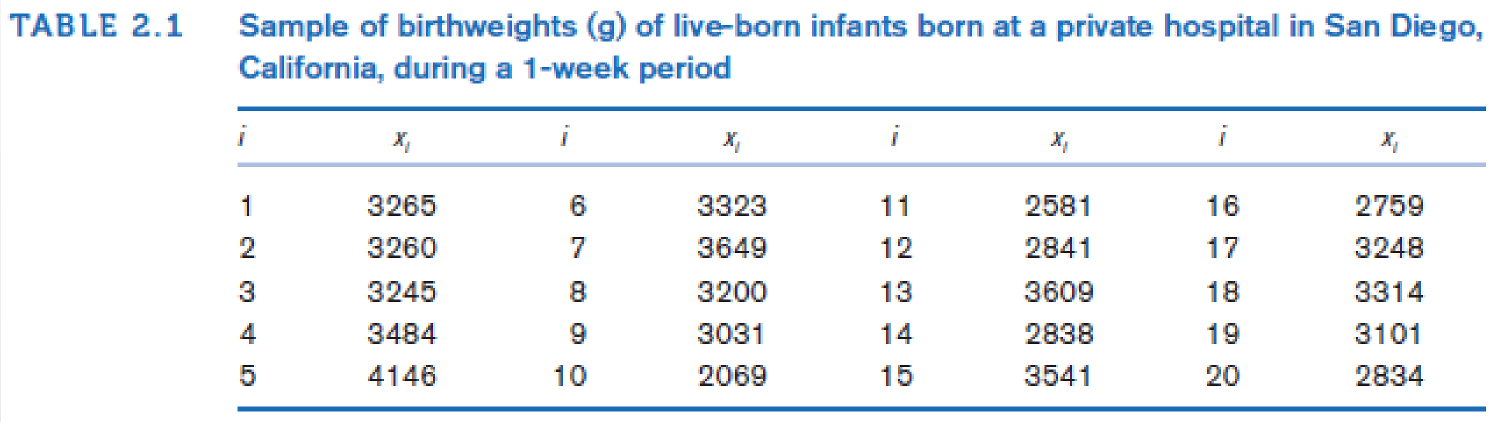
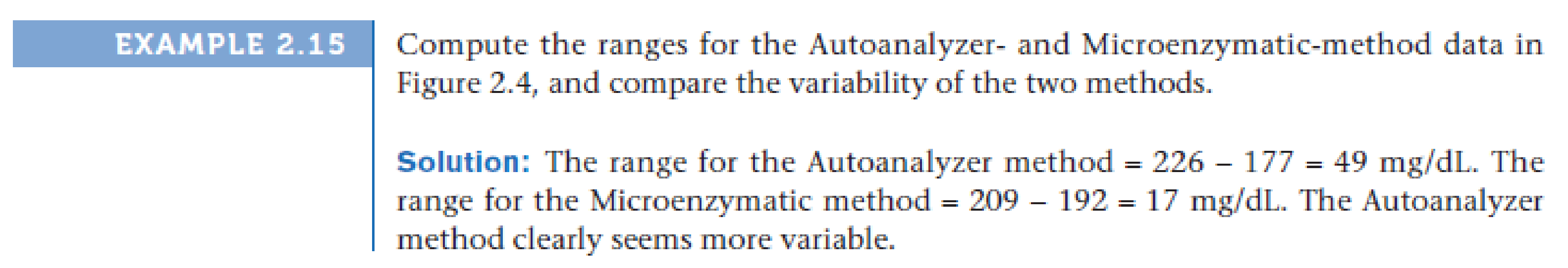


TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

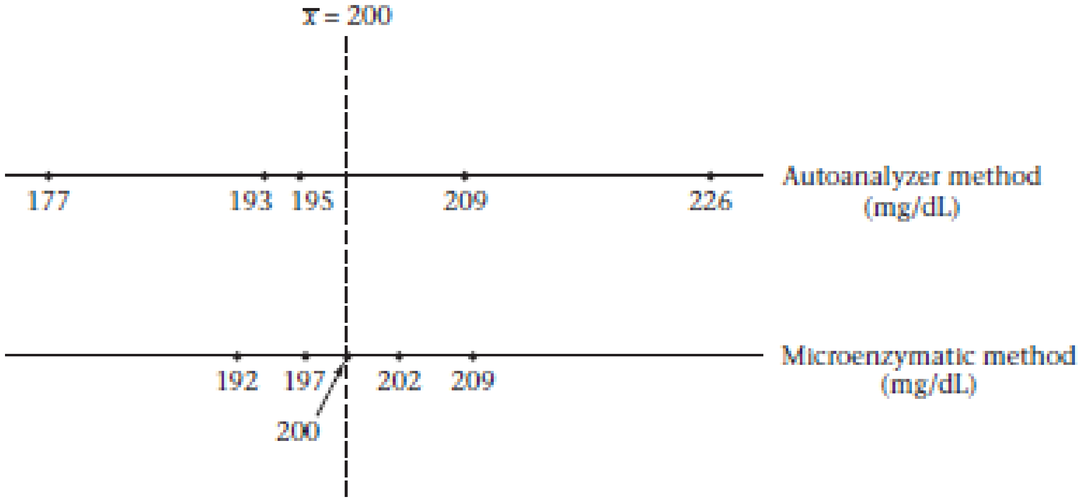
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834



EXAMPLE 2.15 Compute the ranges for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4, and compare the variability of the two methods.

Solution: The range for the Autoanalyzer method = 226 − 177 = 49 mg/dL. The range for the Microenzymatic method = 209 − 192 = 17 mg/dL. The Autoanalyzer method clearly seems more variable.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



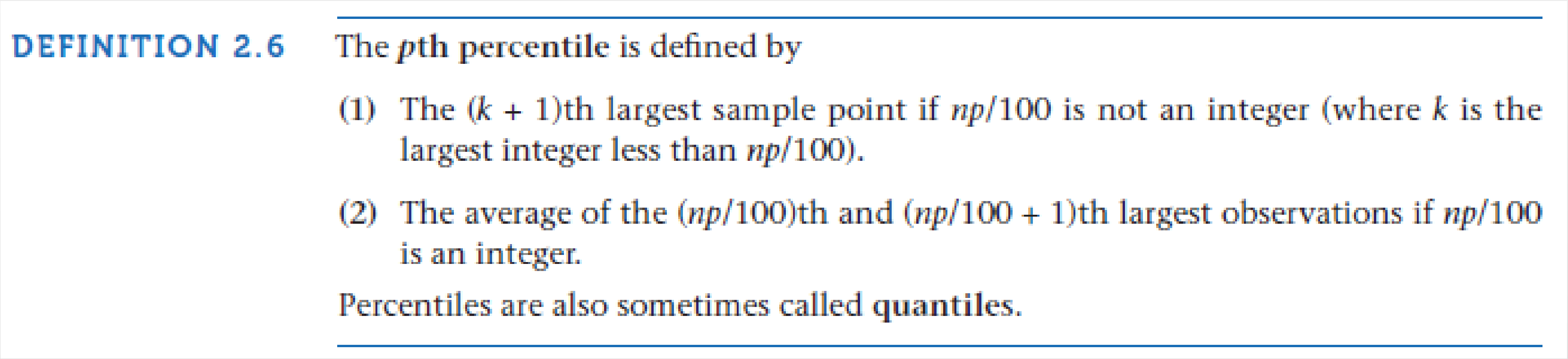
Quantiles (Percentiles)

Shortcomings: eksiklikler

- Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of quantiles or percentiles.
- A procedure for obtaining the p^{th} percentile of a data set of size n is as follows:
- Step1: Arrange the data in ascending (increasing) order
- Step 2: Compute an index i as follows   i=p.n/100
- Step3:
  - If i is an integer, the p^{th} percentile is the average of the i^{th} and $(i +1)^{th}$ smallest data values
  - If i is not an integer then round i up to the nearest integer and take the value at that postion

Quantiles (Percentiles)

- The median, being the 50th percentile, is a special case of a quantile.



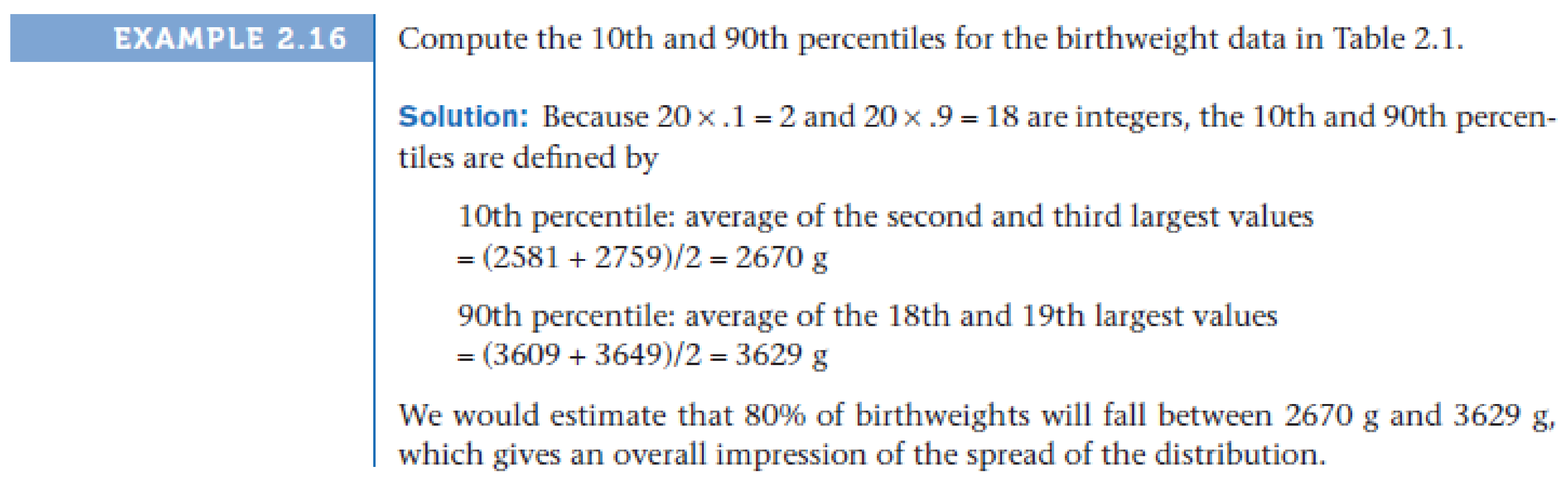
DEFINITION 2.6 The ***p*th percentile** is defined by

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$).

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer.

Percentiles are also sometimes called **quantiles**.

- The spread of a distribution can be characterized by specifying several percentiles.
- For example, the 10th and 90th percentiles are often used to characterize spread.
- Percentiles have the advantage over the range of being less sensitive to outliers and of not being greatly affected by the sample size (*n*).



EXAMPLE 2.16 Compute the 10th and 90th percentiles for the birthweight data in Table 2.1.

Solution: Because $20 \times .1 = 2$ and $20 \times .9 = 18$ are integers, the 10th and 90th percentiles are defined by

10th percentile: average of the second and third largest values
$= (2581 + 2759)/2 = 2670$ g

90th percentile: average of the 18th and 19th largest values
$= (3609 + 3649)/2 = 3629$ g

We would estimate that 80% of birthweights will fall between 2670 g and 3629 g, which gives an overall impression of the spread of the distribution.

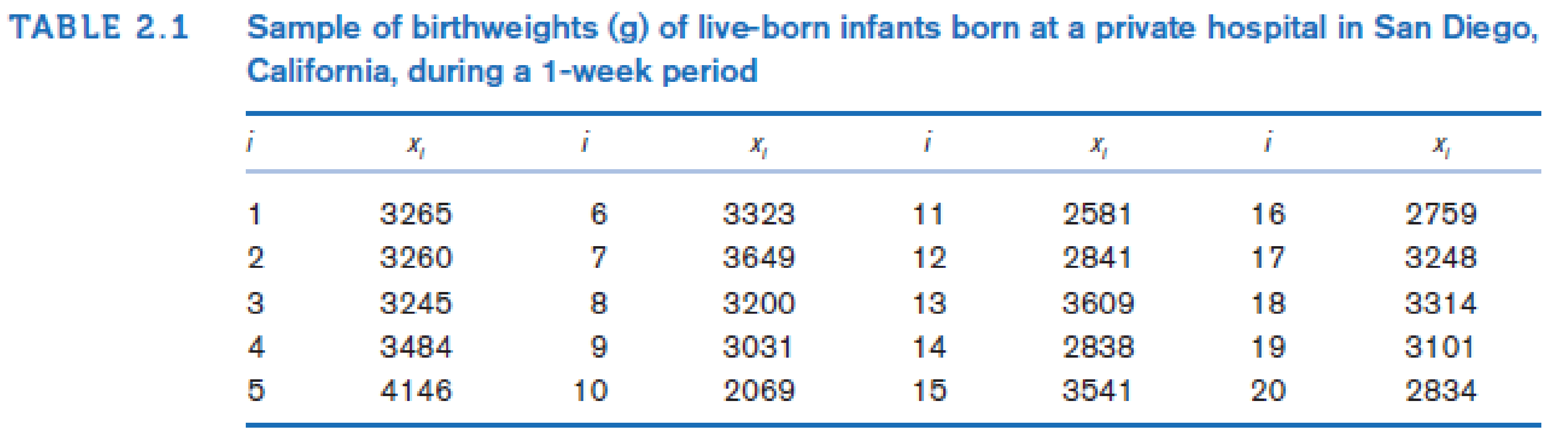
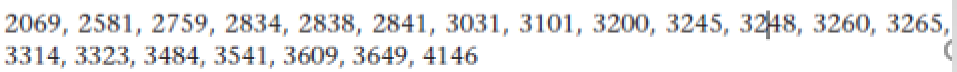


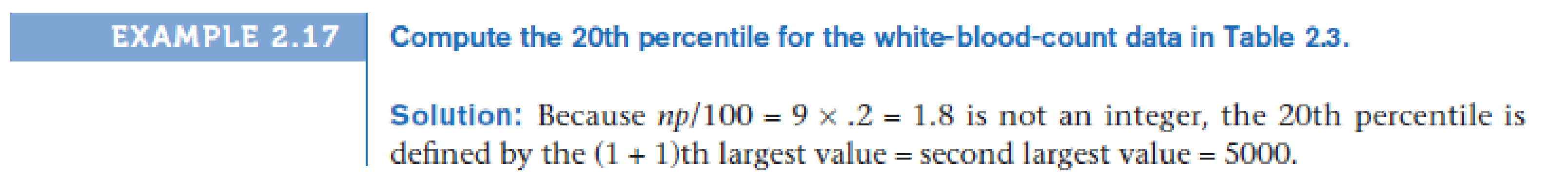
TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834



2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265,
3314, 3323, 3484, 3541, 3609, 3649, 4146

do not forget.



EXAMPLE 2.17 Compute the 20th percentile for the white-blood-count data in Table 2.3.

Solution: Because $np/100 = 9 \times .2 = 1.8$ is not an integer, the 20th percentile is defined by the $(1 + 1)$th largest value = second largest value = 5000.

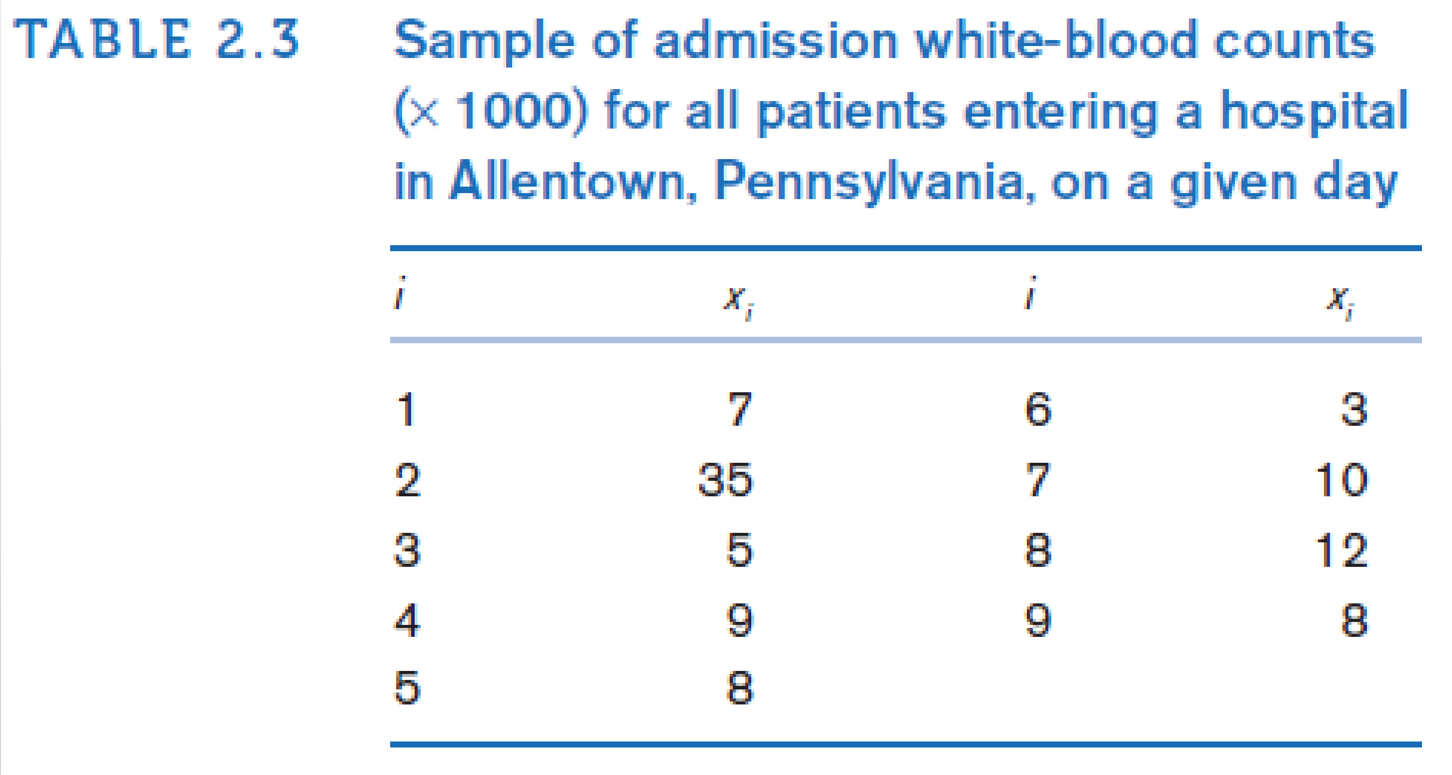


TABLE 2.3 Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

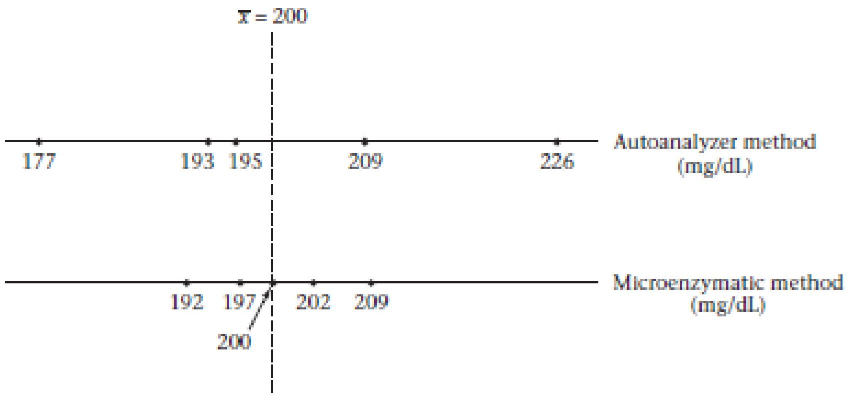
i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Percentiles not only give locate the center of a distribution but also other locations in a distribution

The Variance and Standard Deviation

- The main difference between the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 is that the Microenzymatic-method values are closer to the center of the sample than the Autoanalyzer-method values.
- If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean is needed.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



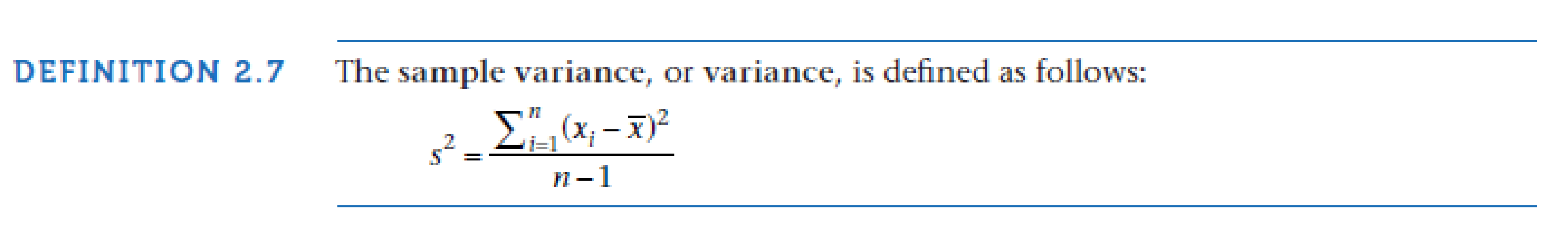
The Variance and Standard Deviation Example

- Biomedical Engineering age

- Electronics engineer ig age

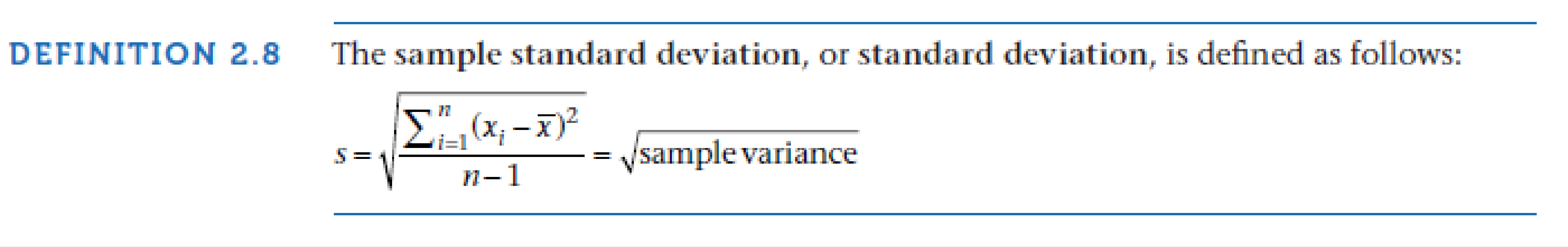
Variance and Standard Deviation

- The most important measures of dispersion are the variance and its square root, the standard deviation.
- Since the variance is just the square of the standard deviation, these quantities contain essentially the same information, just on different scales.



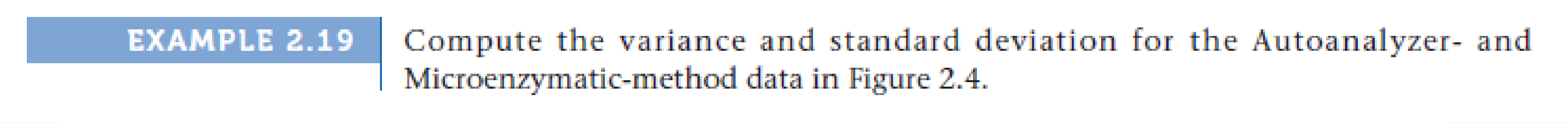
DEFINITION 2.7 The sample variance, or variance, is defined as follows:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

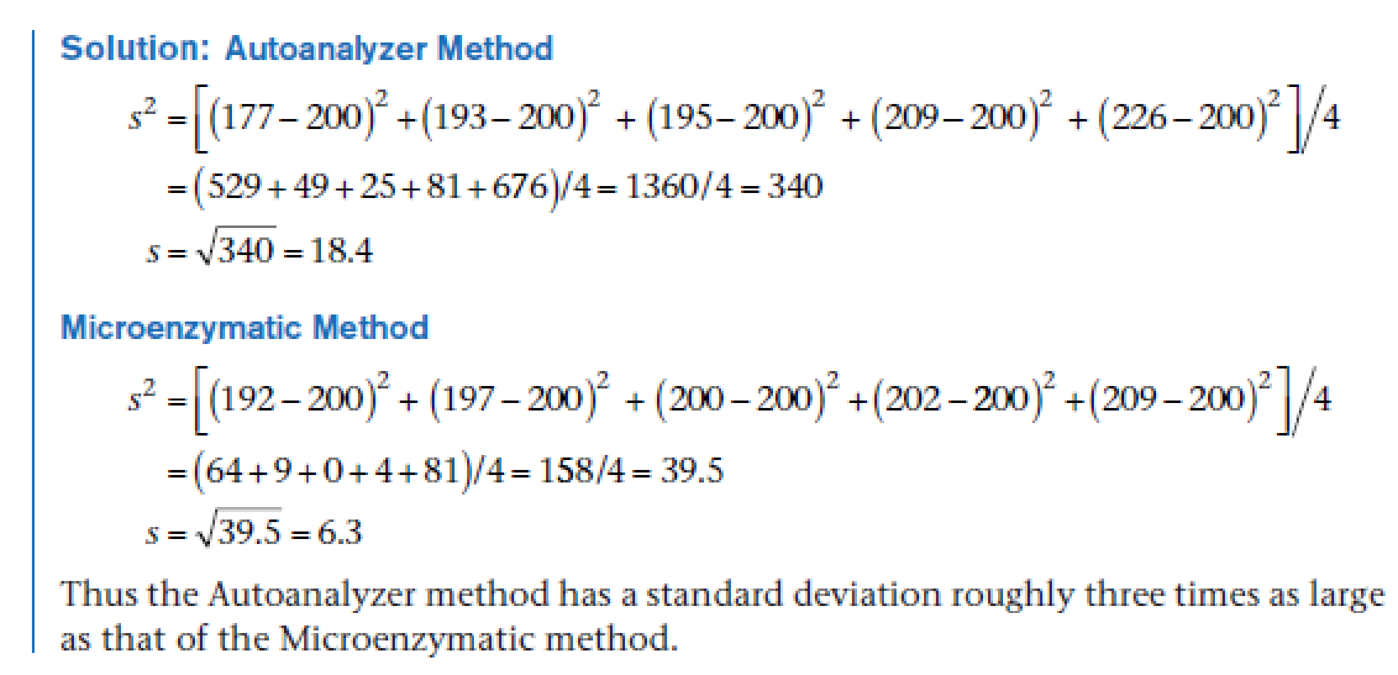


DEFINITION 2.8 The sample standard deviation, or standard deviation, is defined as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$



EXAMPLE 2.19 Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.



Solution: Autoanalyzer Method

$$s^2 = \left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]/4$$

$$= (529+49+25+81+676)/4 = 1360/4 = 340$$

$$s = \sqrt{340} = 18.4$$

Microenzymatic Method

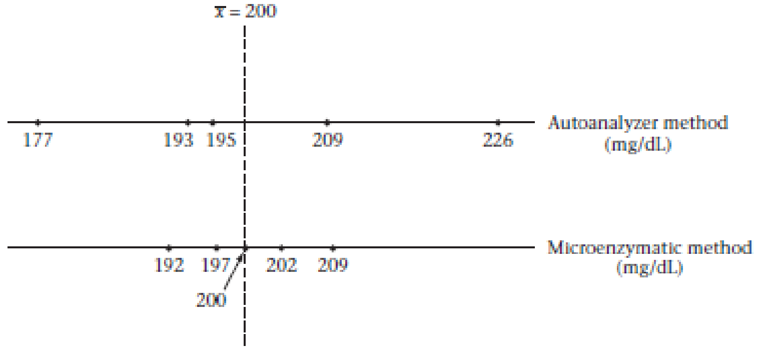
$$s^2 = \left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]/4$$

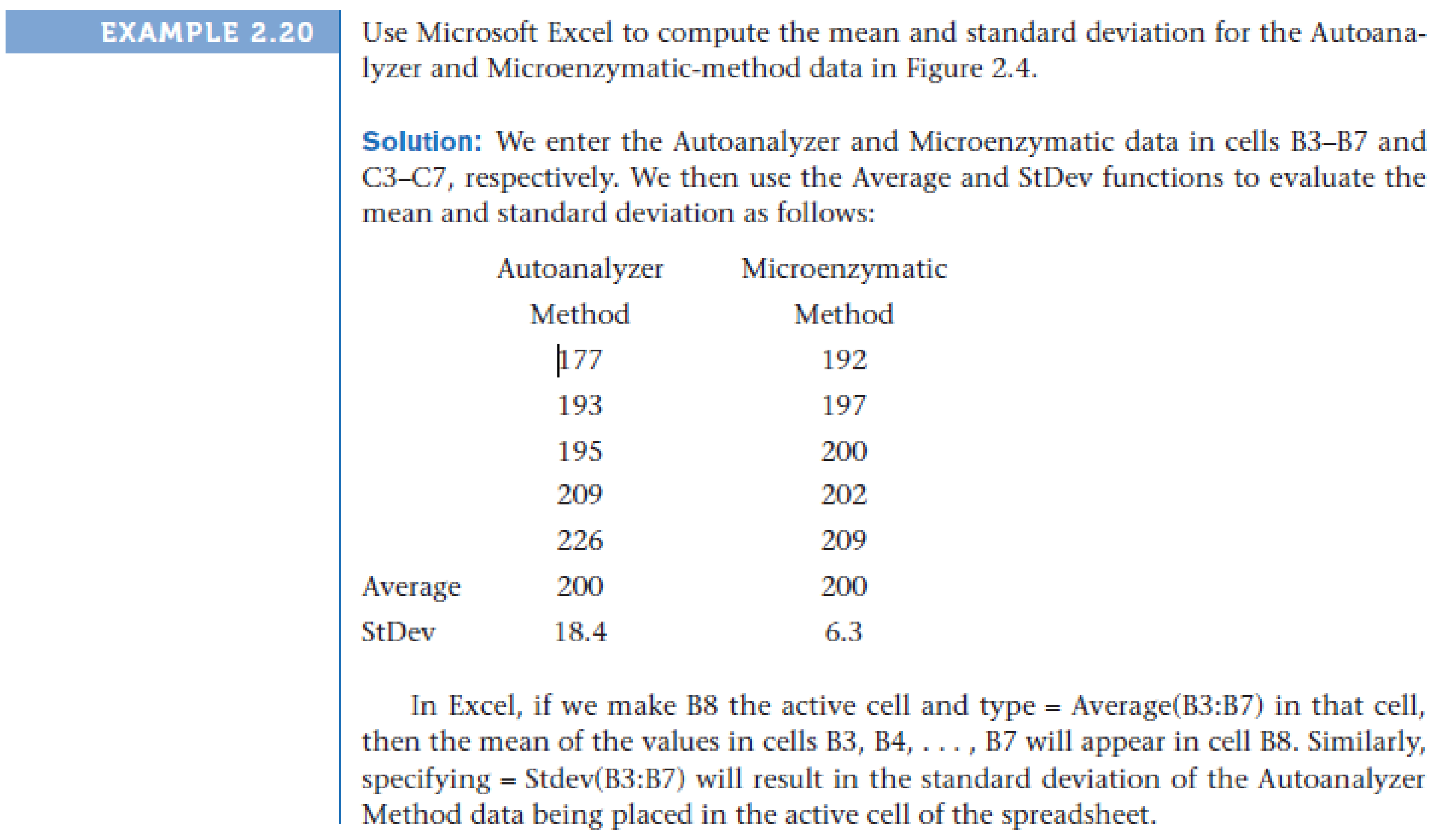
$$= (64+9+0+4+81)/4 = 158/4 = 39.5$$

$$s = \sqrt{39.5} = 6.3$$

Thus the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



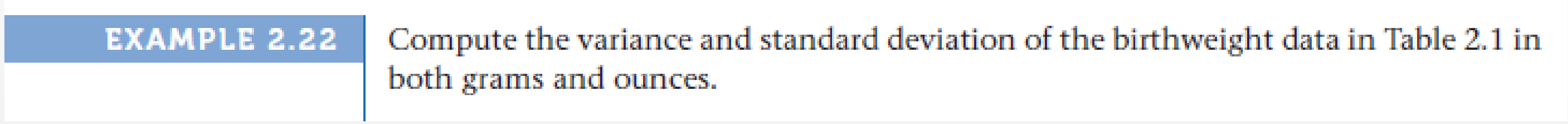


EXAMPLE 2.20 Use Microsoft Excel to compute the mean and standard deviation for the Autoanalyzer and Microenzymatic-method data in Figure 2.4.

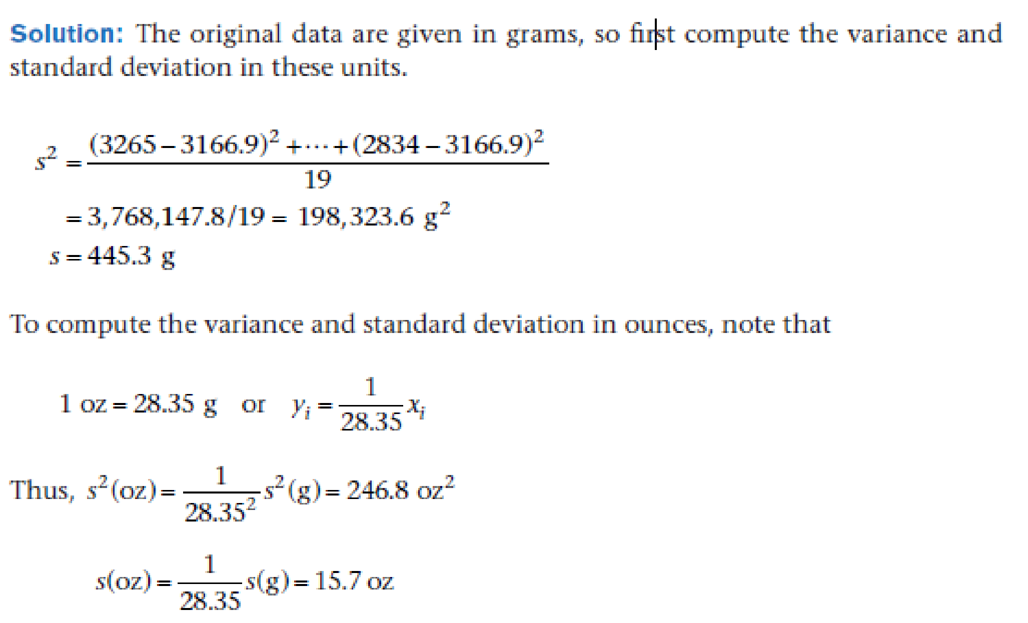
Solution: We enter the Autoanalyzer and Microenzymatic data in cells B3–B7 and C3–C7, respectively. We then use the Average and StDev functions to evaluate the mean and standard deviation as follows:

	Autoanalyzer	Microenzymatic
	Method	Method
	177	192
	193	197
	195	200
	209	202
	226	209
Average	200	200
StDev	18.4	6.3

In Excel, if we make B8 the active cell and type = Average(B3:B7) in that cell, then the mean of the values in cells B3, B4, . . . , B7 will appear in cell B8. Similarly, specifying = Stdev(B3:B7) will result in the standard deviation of the Autoanalyzer Method data being placed in the active cell of the spreadsheet.



EXAMPLE 2.22 Compute the variance and standard deviation of the birthweight data in Table 2.1 in both grams and ounces.



Solution: The original data are given in grams, so first compute the variance and standard deviation in these units.

$$s^2 = \frac{(3265-3166.9)^2+\cdots+(2834-3166.9)^2}{19}$$

$$= 3{,}768{,}147.8/19 = 198{,}323.6 \text{ g}^2$$

$$s = 445.3 \text{ g}$$

To compute the variance and standard deviation in ounces, note that

$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35}x_i$$

Thus, $s^2(\text{oz}) = \frac{1}{28.35^2}s^2(\text{g}) = 246.8 \text{ oz}^2$

$$s(\text{oz}) = \frac{1}{28.35}s(\text{g}) = 15.7 \text{ oz}$$

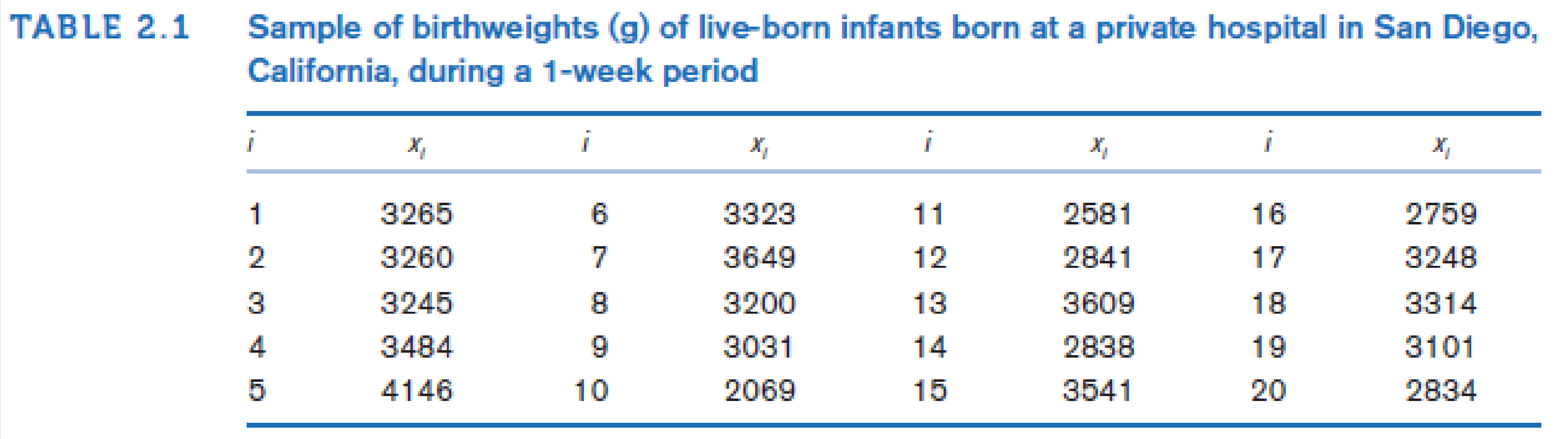
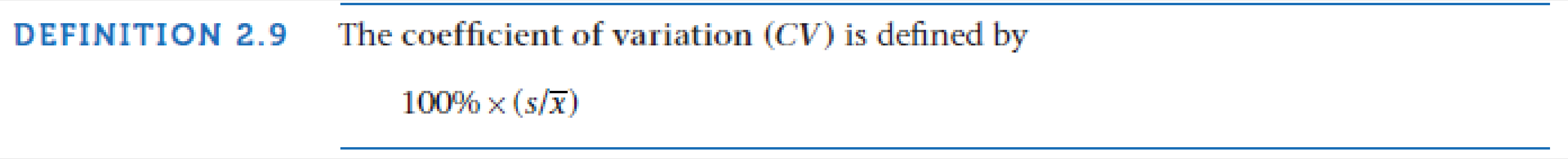


TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

The Coefficient of Variation

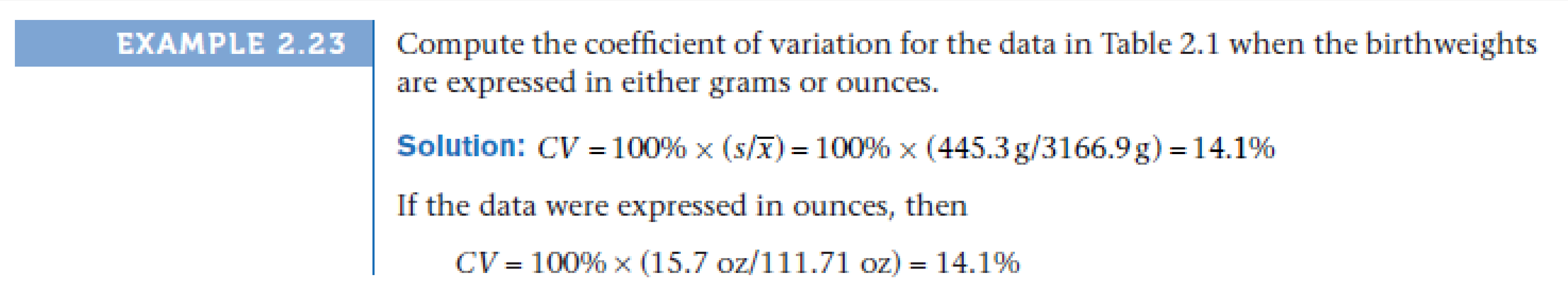
- It is useful to relate the arithmetic mean and the standard deviation to each other because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 versus if it is 1000.
- A special measure, the coefficient of variation, is often used for this purpose.



DEFINITION 2.9 The coefficient of variation (CV) is defined by

$$100\% \times (s/\overline{x})$$

- This measure remains the same regardless of what units are used
- because if the units change by a factor *c*,
- then both the mean and standard deviation change by the factor *c;*
- while the *CV*, which is the ratio between them, remains unchanged.



EXAMPLE 2.23 Compute the coefficient of variation for the data in Table 2.1 when the birthweights are expressed in either grams or ounces.

Solution: $CV = 100\% \times (s/\overline{x}) = 100\% \times (445.3\,\text{g}/3166.9\,\text{g}) = 14.1\%$

If the data were expressed in ounces, then

$$CV = 100\% \times (15.7\text{ oz}/111.71\text{ oz}) = 14.1\%$$

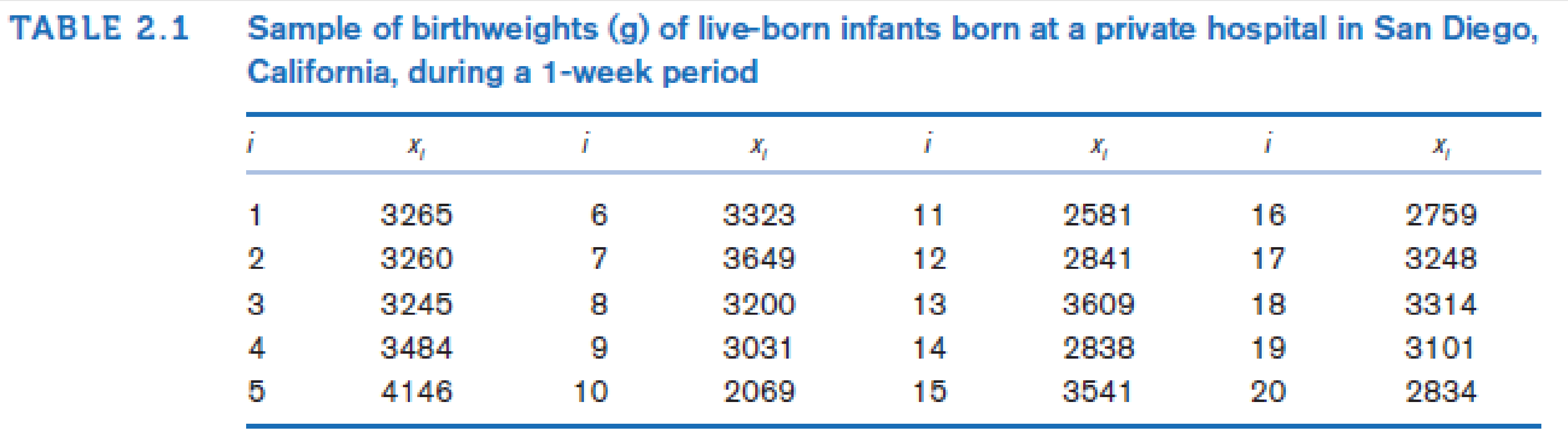


TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- The *CV* is most useful in comparing the variability of several different samples, each with different arithmetic means.

- This is because a higher variability is usually expected when the mean increases, and the *CV* is a measure that accounts for this variability.

Grouped Data

- Sometimes the sample size is too large to display all the raw data.
- Also, data are frequently collected in grouped
- Although a set of observation can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data.

Grouped Data

- Consider the data set in Table 2.9, which represents the birthweights from 100 consecutive deliveries at a Boston hospital.

- Suppose we wish to display these data for publication purposes.
- How can we do this?
- The simplest way to display the data is to generate a frequency distribution using a statistical package

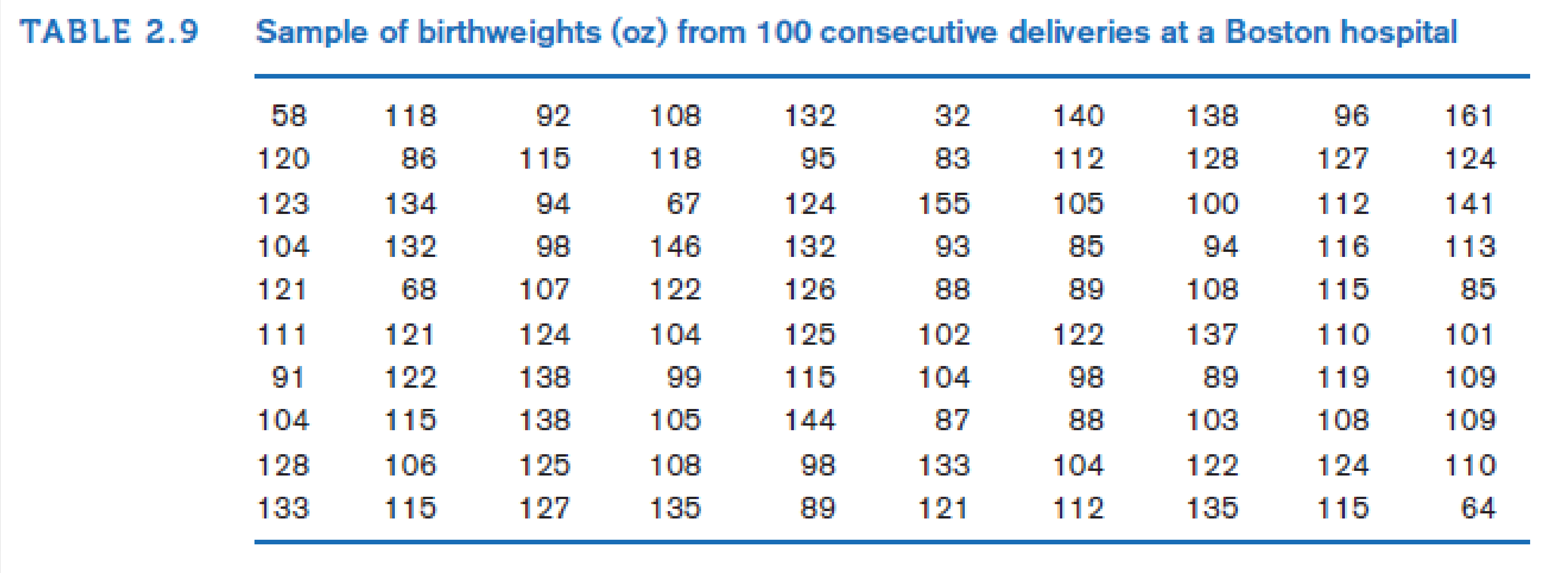


TABLE 2.9 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

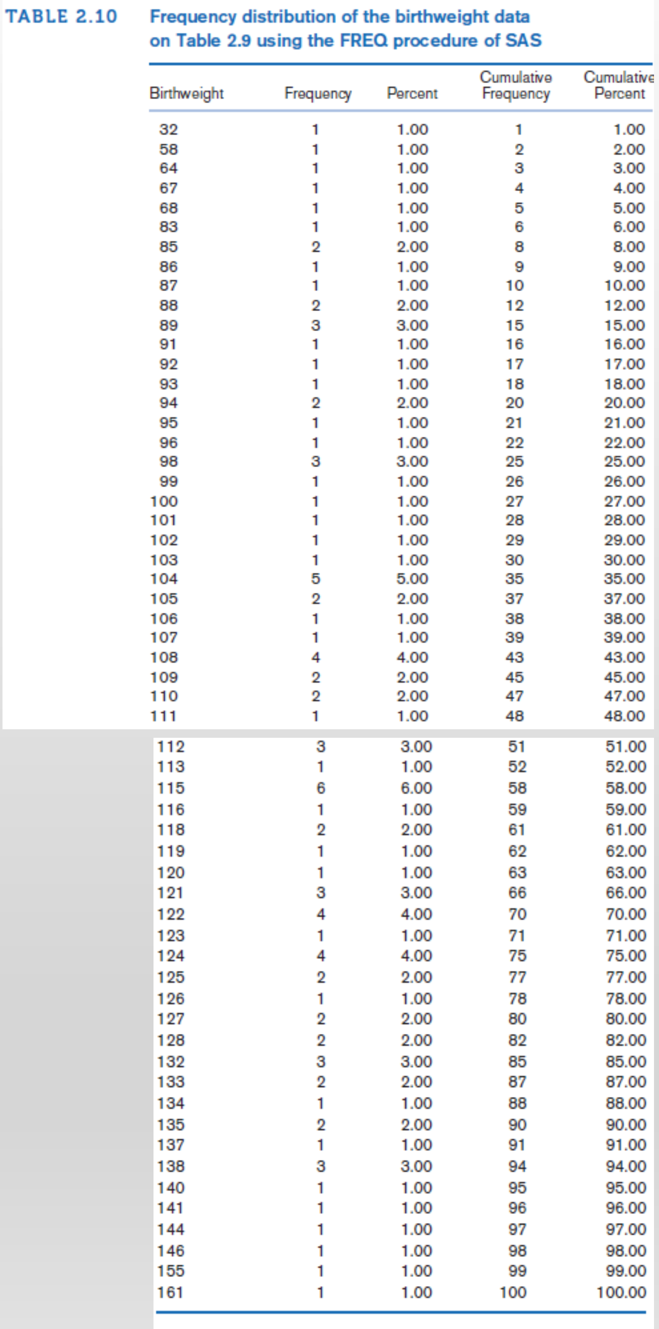
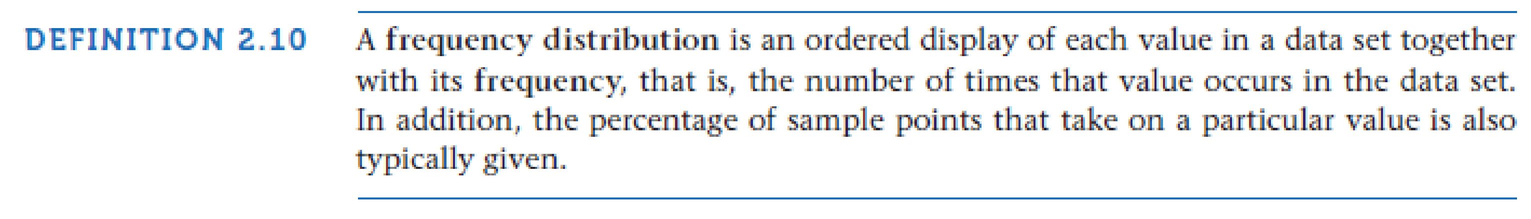


TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00



DEFINITION 2.10 A **frequency distribution** is an ordered display of each value in a data set together with its frequency, that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given.

- A frequency distribution of the sample of 100 birthweights in Table 2.9, generated is displayed in Table 2.10.
- frequency procedure provides the
  - Frequency,
  - relative frequency (Percent),
  - Cumulative Frequency,
  - Cumulative Percent for each birthweight present in the sample.

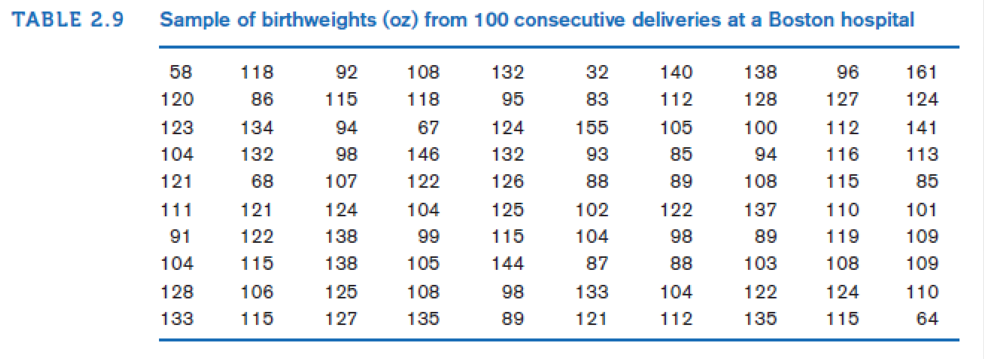


TABLE 2.9 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

- For any particular birthweight *b*,
- The Cumulative Frequency is the number of birthweights in the sample that are less than or equal to *b.*
- The Percent = 100 × Frequency/*n*,
- Cumulative Percent = 100 × Cumulative Frequency/*n*
- Cumulative Percent = the percentage of birthweights less than or equal to *b.*

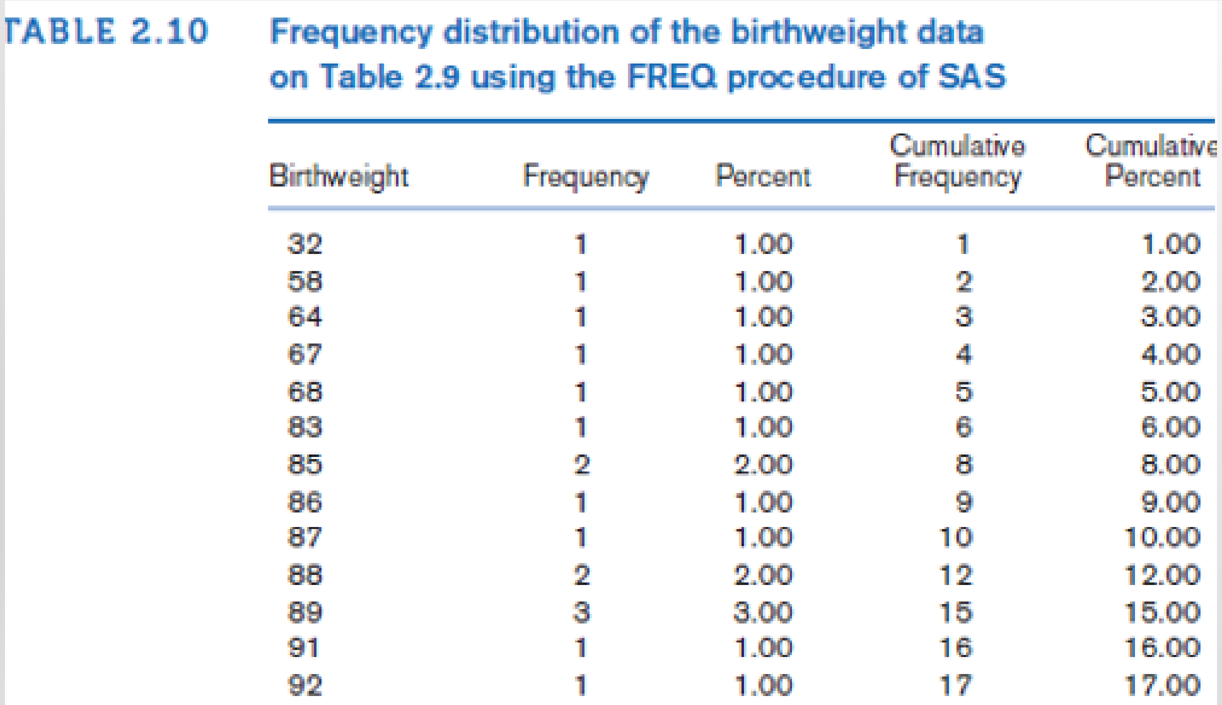


TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00

- 1 oz =28.35 gram

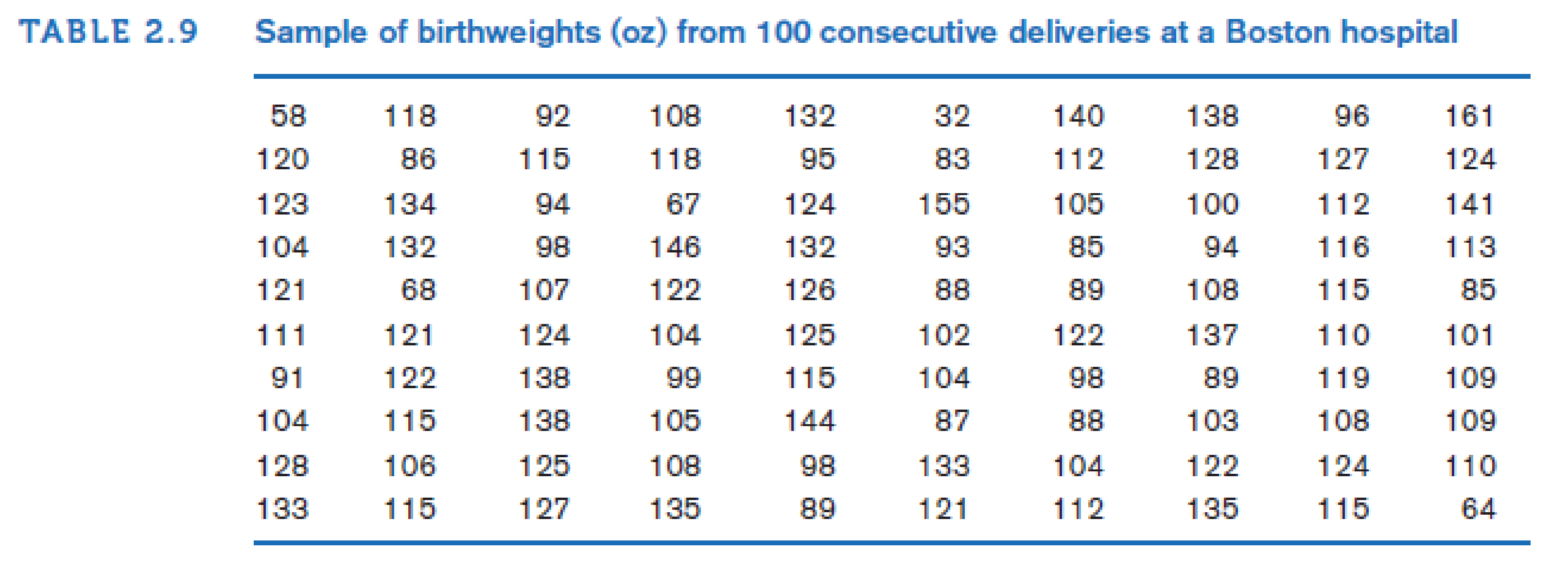
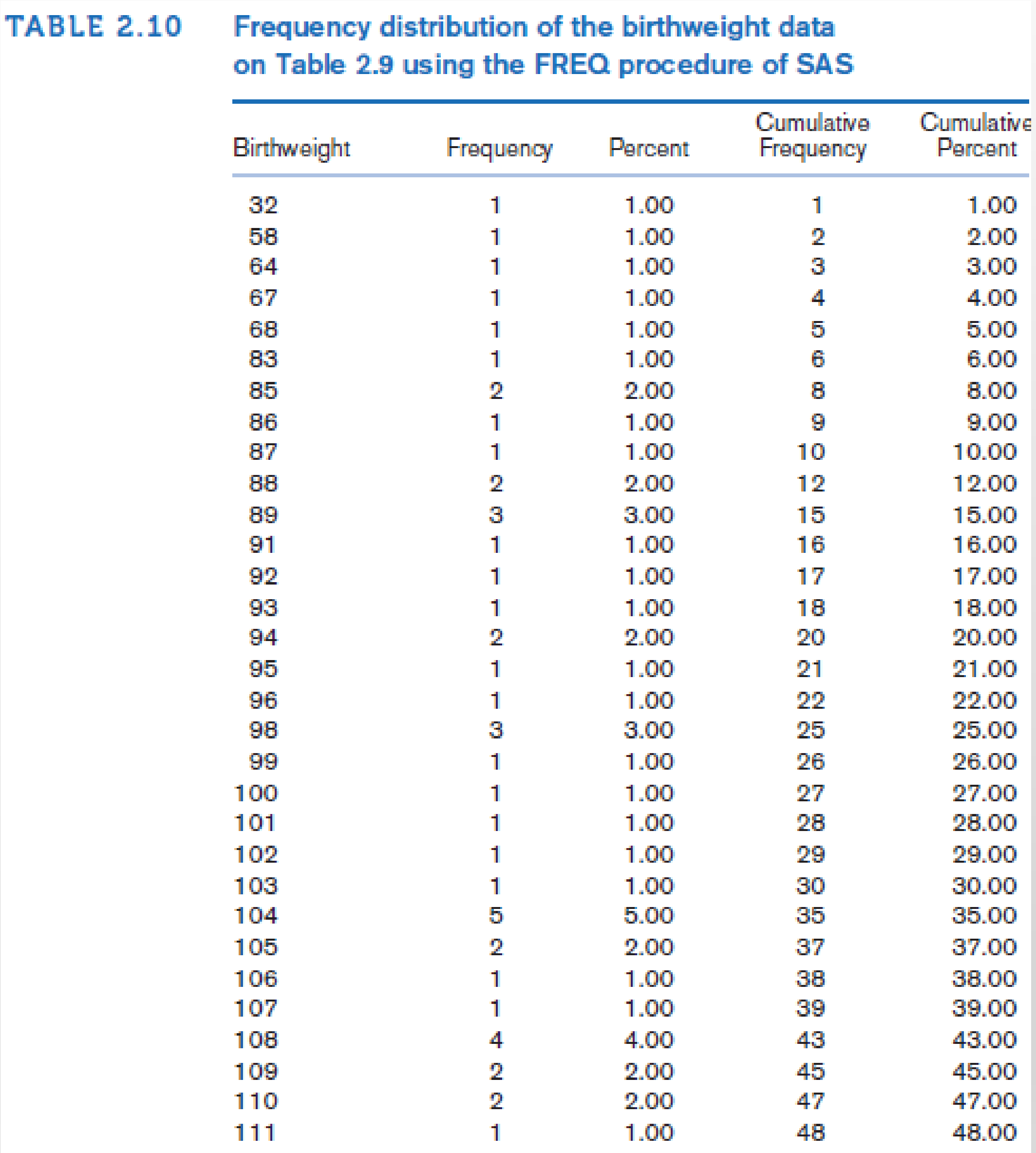


TABLE 2.9 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64



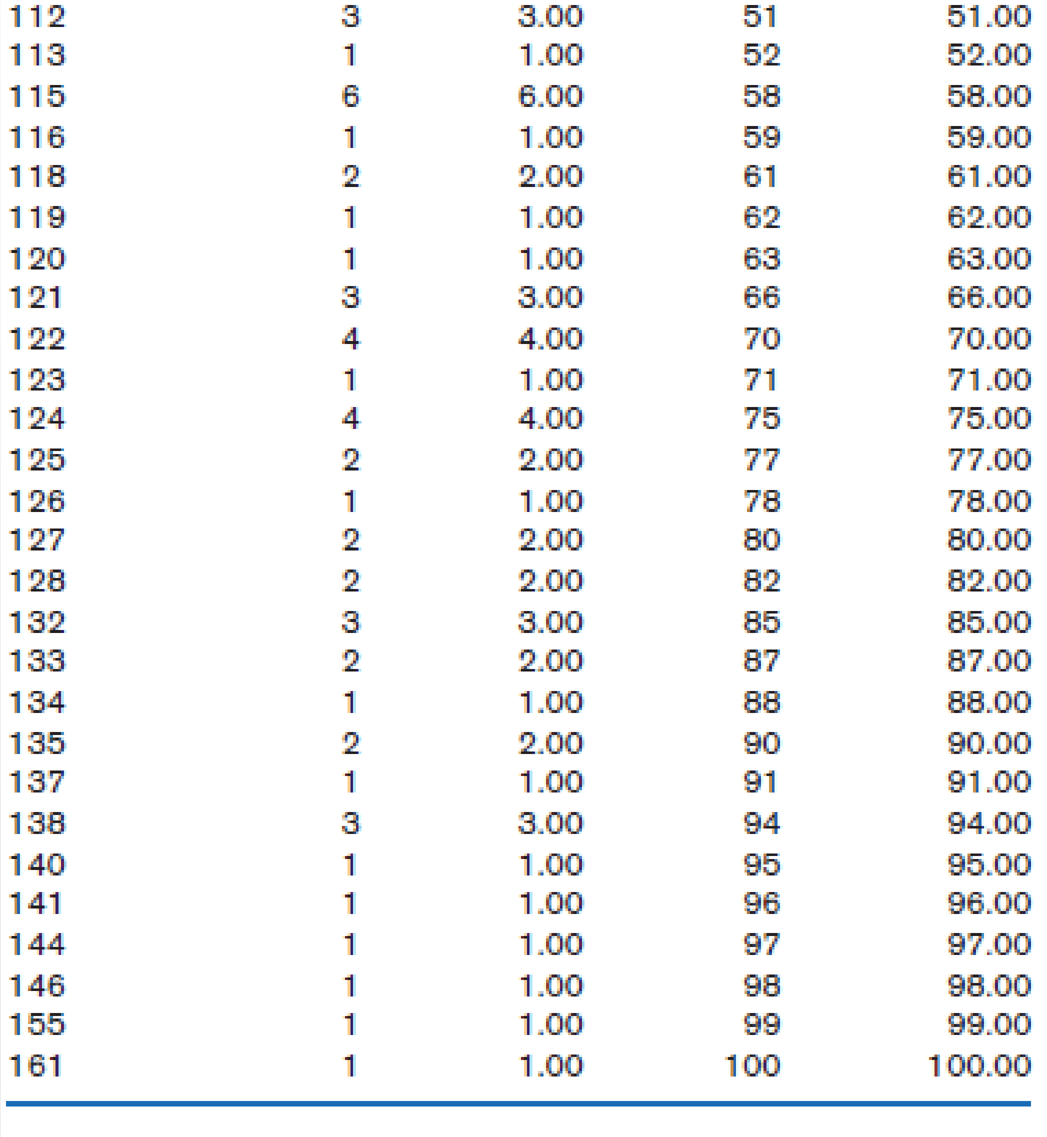


TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00

- The main purpose of grouping is to make the data better understood and look better.
- The people who collect the data can make groupings as they wish.
- They can divide them into as many groups as they want.
- Make sure that the data is understandable when dividing into groups.
- When grouping, the amount of data is very important.
- An example of grouping by data numbers.
- But you do not have to do so.
- you can do as you wish.

if there is 20 data, make 5 groups
if there is 50 data, make 7 groups
if there is 100 data, make 8 groups
if there is 1000 data, make 11 groups

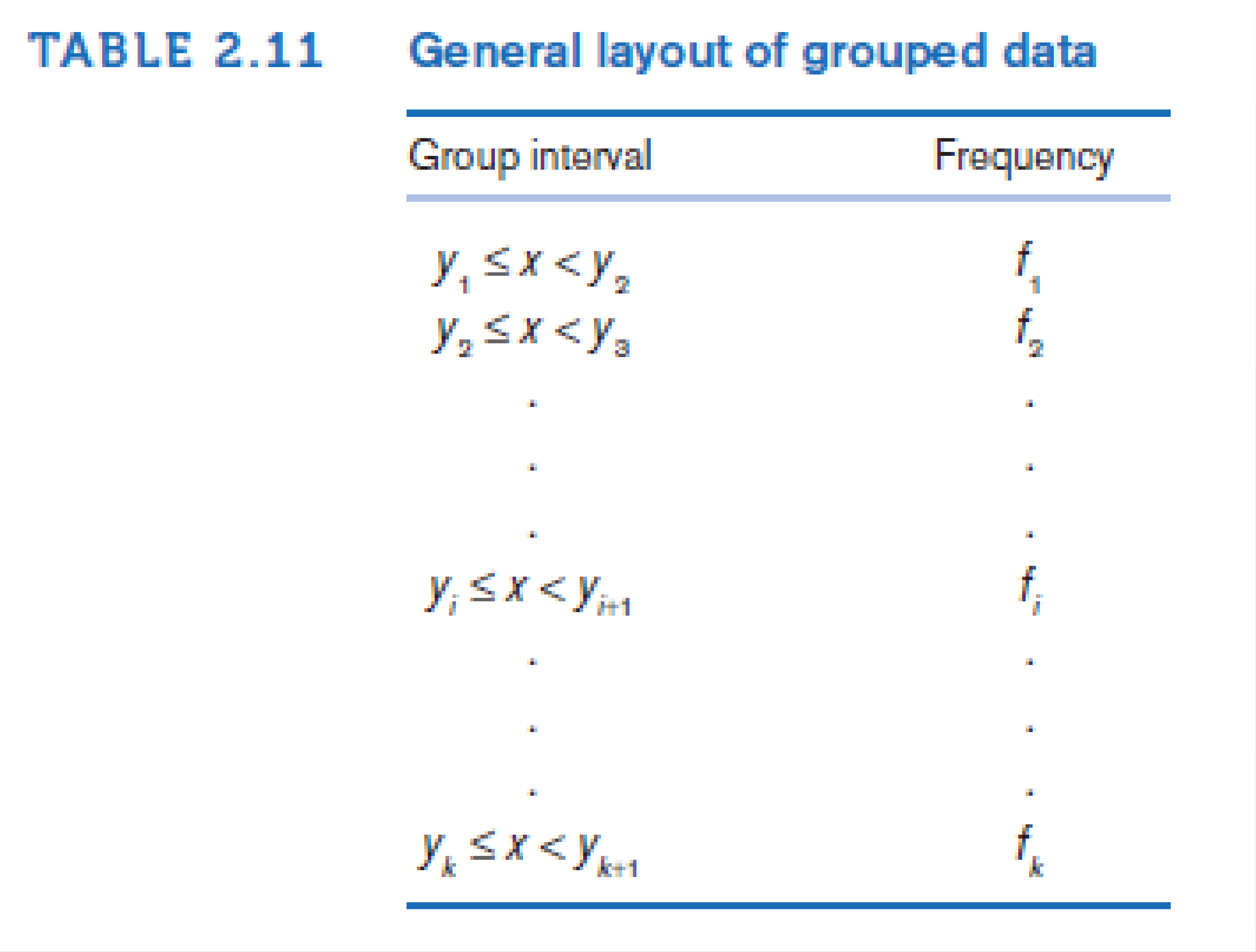


TABLE 2.11 General layout of grouped data

Group interval	Frequency
$y_1 \leq x < y_2$	f_1
$y_2 \leq x < y_3$	f_2
.	.
.	.
.	.
$y_i \leq x < y_{i+1}$	f_i
.	.
.	.
.	.
$y_k \leq x < y_{k+1}$	f_k

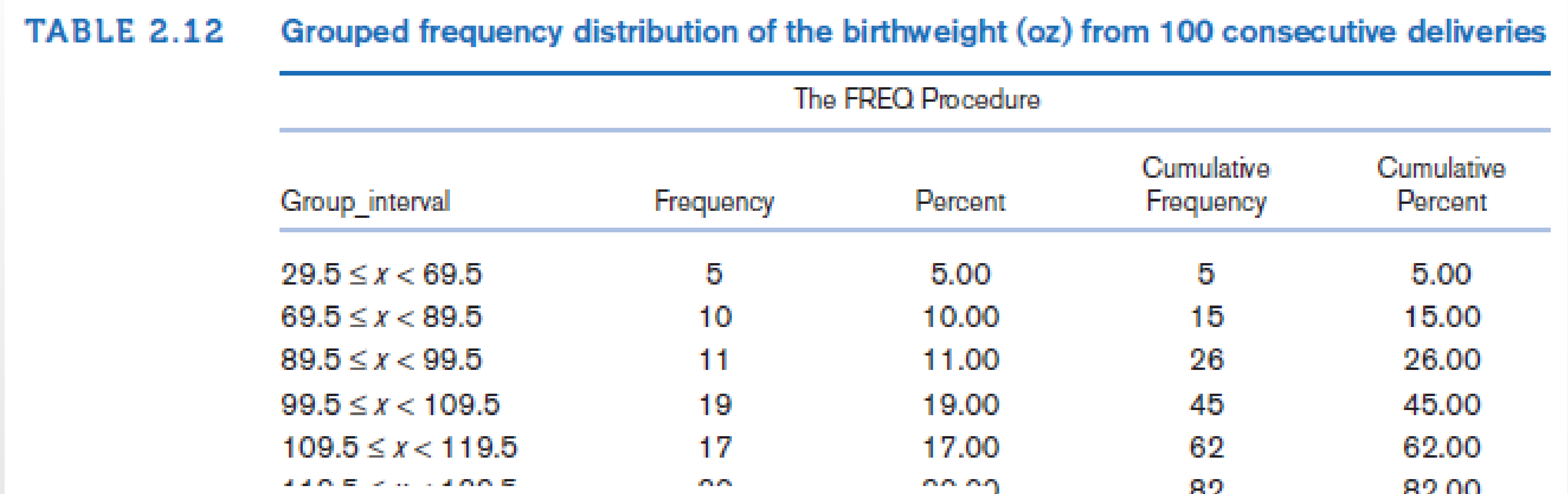


TABLE 2.12 Grouped frequency distribution of the birthweight (oz) from 100 consecutive deliveries

The FREQ Procedure

Group_interval	Frequency	Percent	Cumulative Frequency	Cumulative Percent
29.5 ≤ x < 69.5	5	5.00	5	5.00
69.5 ≤ x < 89.5	10	10.00	15	15.00
89.5 ≤ x < 99.5	11	11.00	26	26.00
99.5 ≤ x < 109.5	19	19.00	45	45.00
109.5 ≤ x < 119.5	17	17.00	62	62.00
119.5 ≤ x < 129.5	20	20.00	82	82.00
			94	94.00
			100	100.00

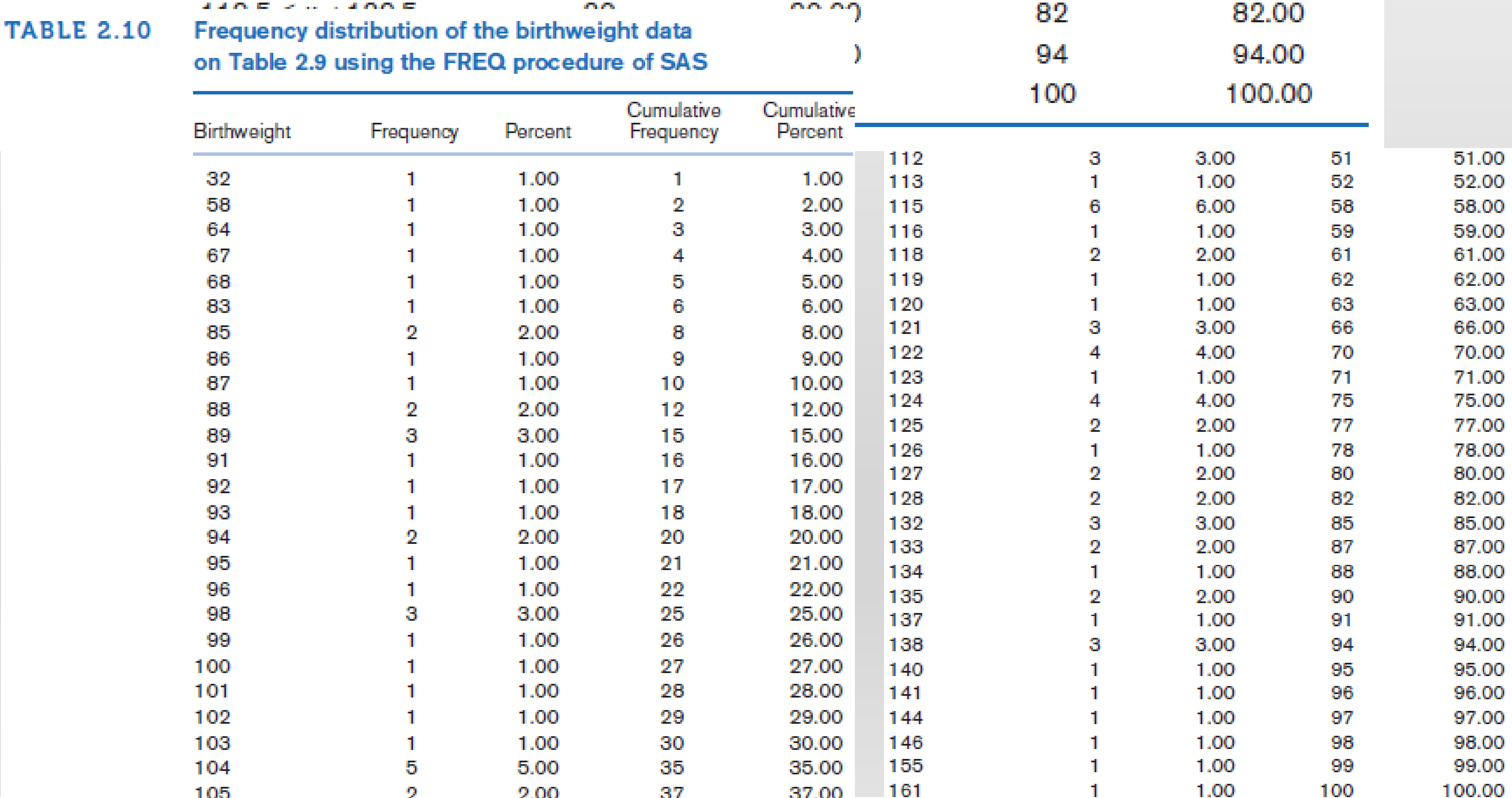
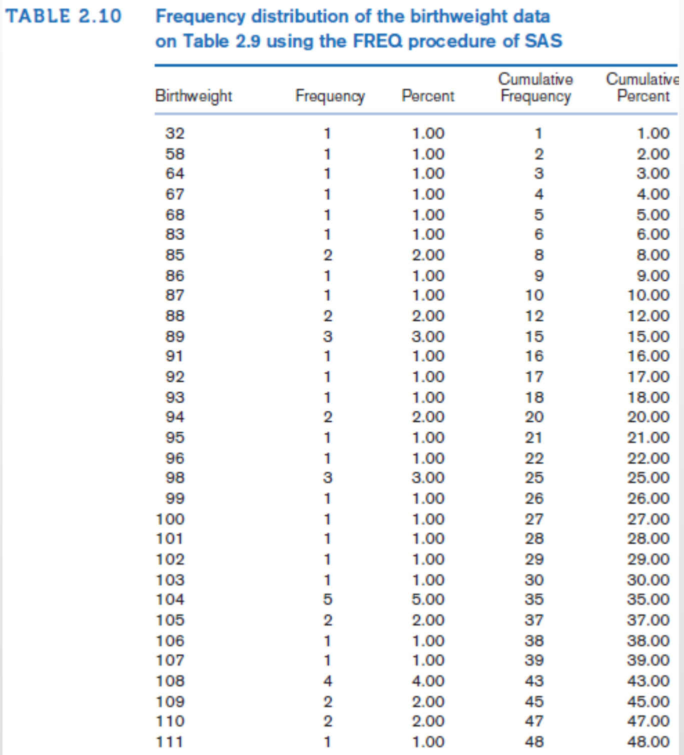


TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00



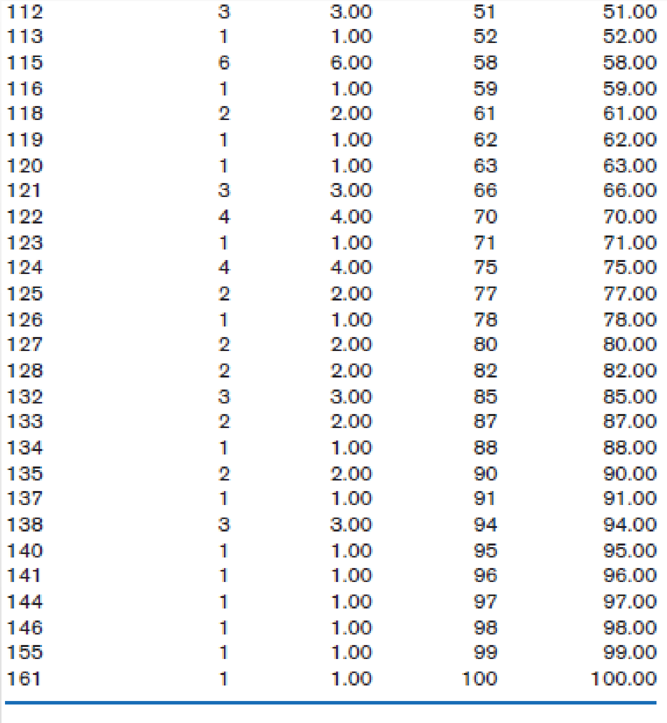


TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00

Please different group

Data

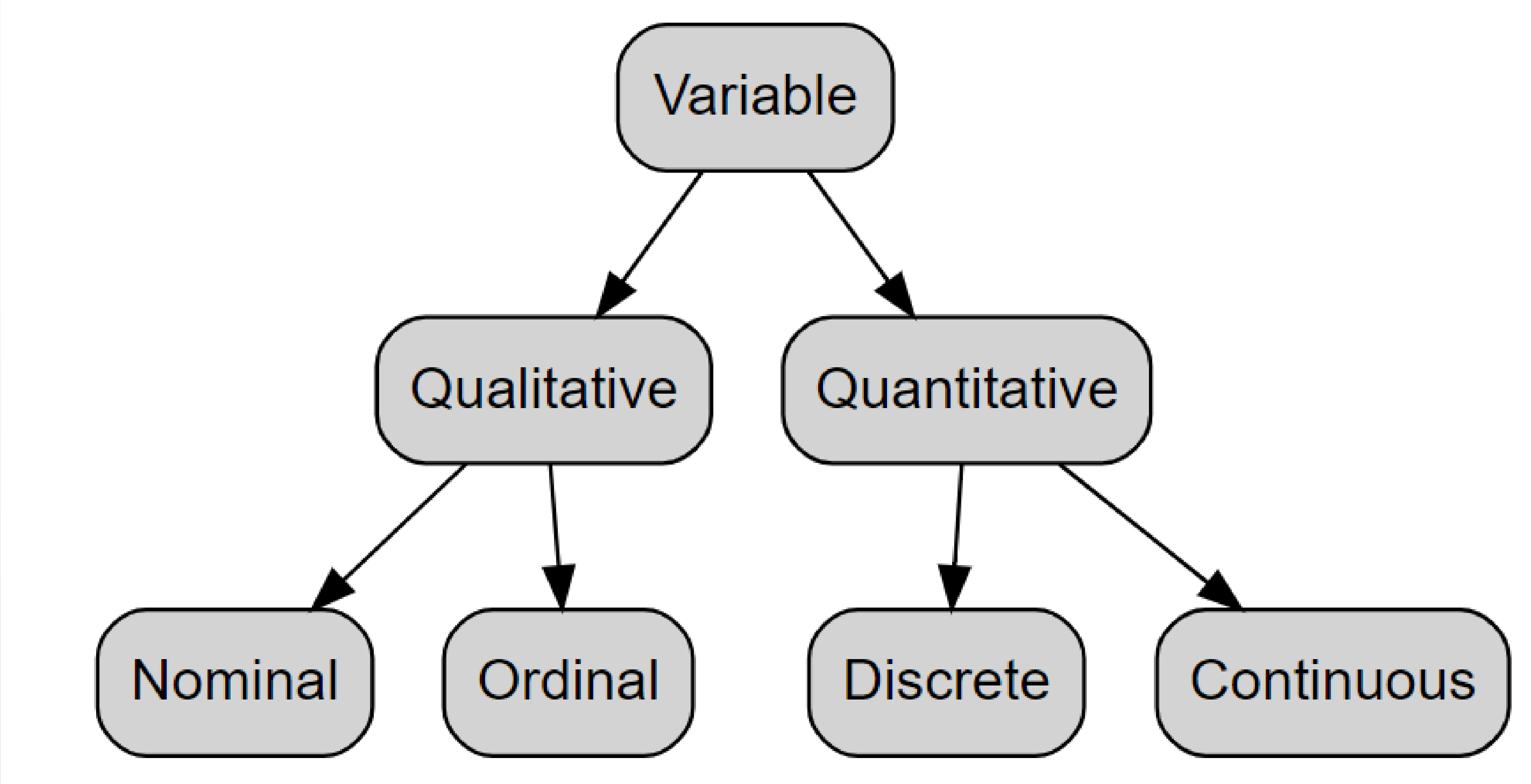
- The raw material of statistics is data
- For our purposes, we may define data as number
- There are two kinds of data
    - Measurement
    - Counting

Ex. When a nurse weighs a patient or takes a patients temperature, a measurement, consist of a number, such as 50kg or 37 oC

Ex. the number of people coming to the hospital during the day

Variables

- In statistics, a **variable** has two defining characteristics:
  - A variable is an attribute that describes a person, place, thing, or idea.
  - The value of the variable can "vary" from one entity to another.
- For example,
  - a person's *hair color* is a potential variable, which could have the value of "blond" for one person and "brunette" for another.

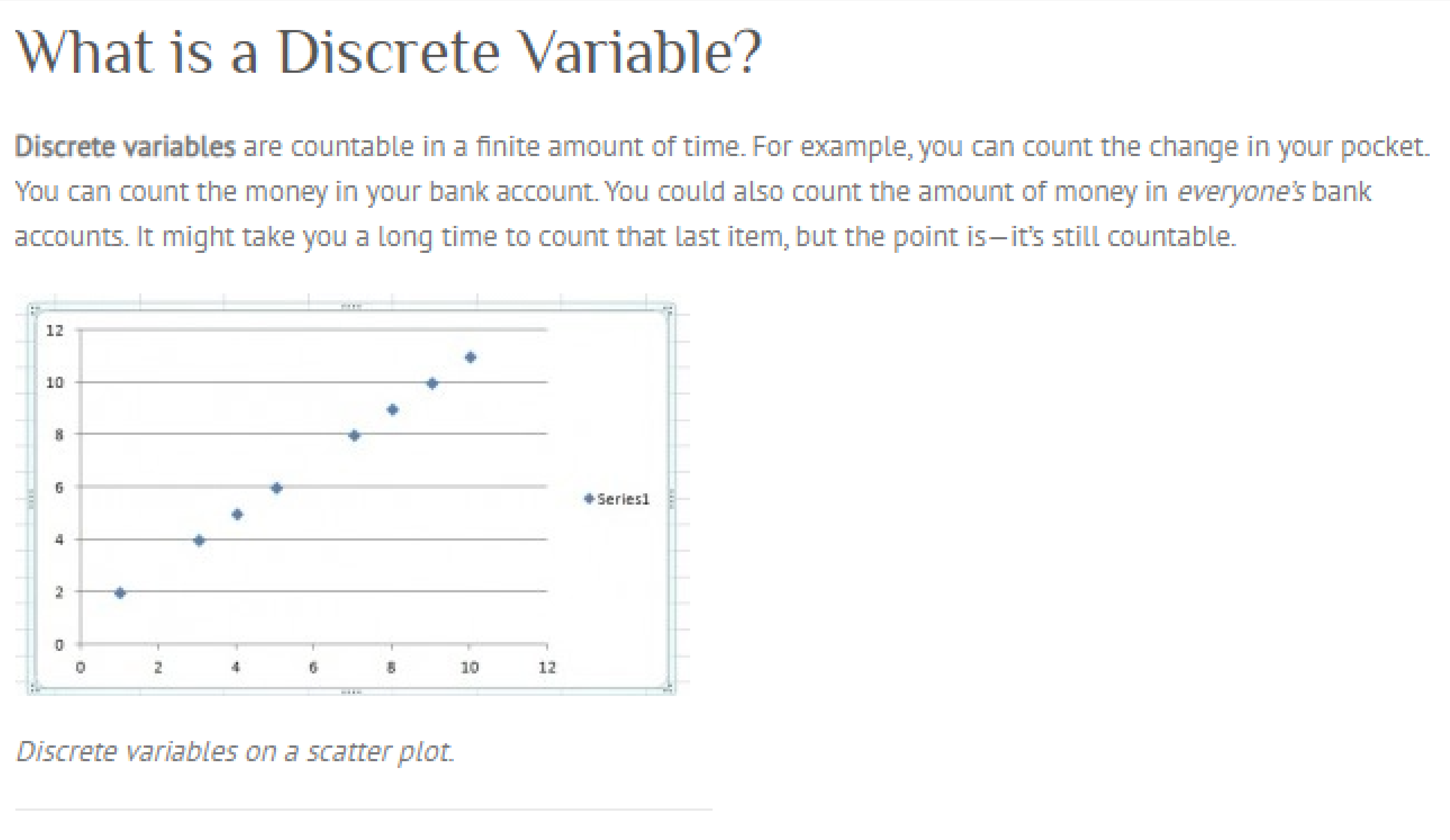

Variable
Qualitative
Quantitative
Nominal
Ordinal
Discrete
Continuous

Quantitative Variables

- A **quantitative** variable is a variable that reflects a notion of **magnitude**, that is, if the values it can take are **numbers**.
- A quantitative variable represents thus a measure and is numerical.
- Quantitative variables are divided into two types:
  - **discrete**
  - **continuous**.

Quantitative discrete variables

- **Quantitative discrete** variables are variables for which the values it can take are **countable** and have a **finite number of possibilities**.
- The values are often integers.
  - Number of children per family
  - Number of students in a class
  - Number of citizens of a country
  - Whatever the number of children in a family, it will never be 3.58 or 7.912 so the number of possibilities is a finite number and thus countable.



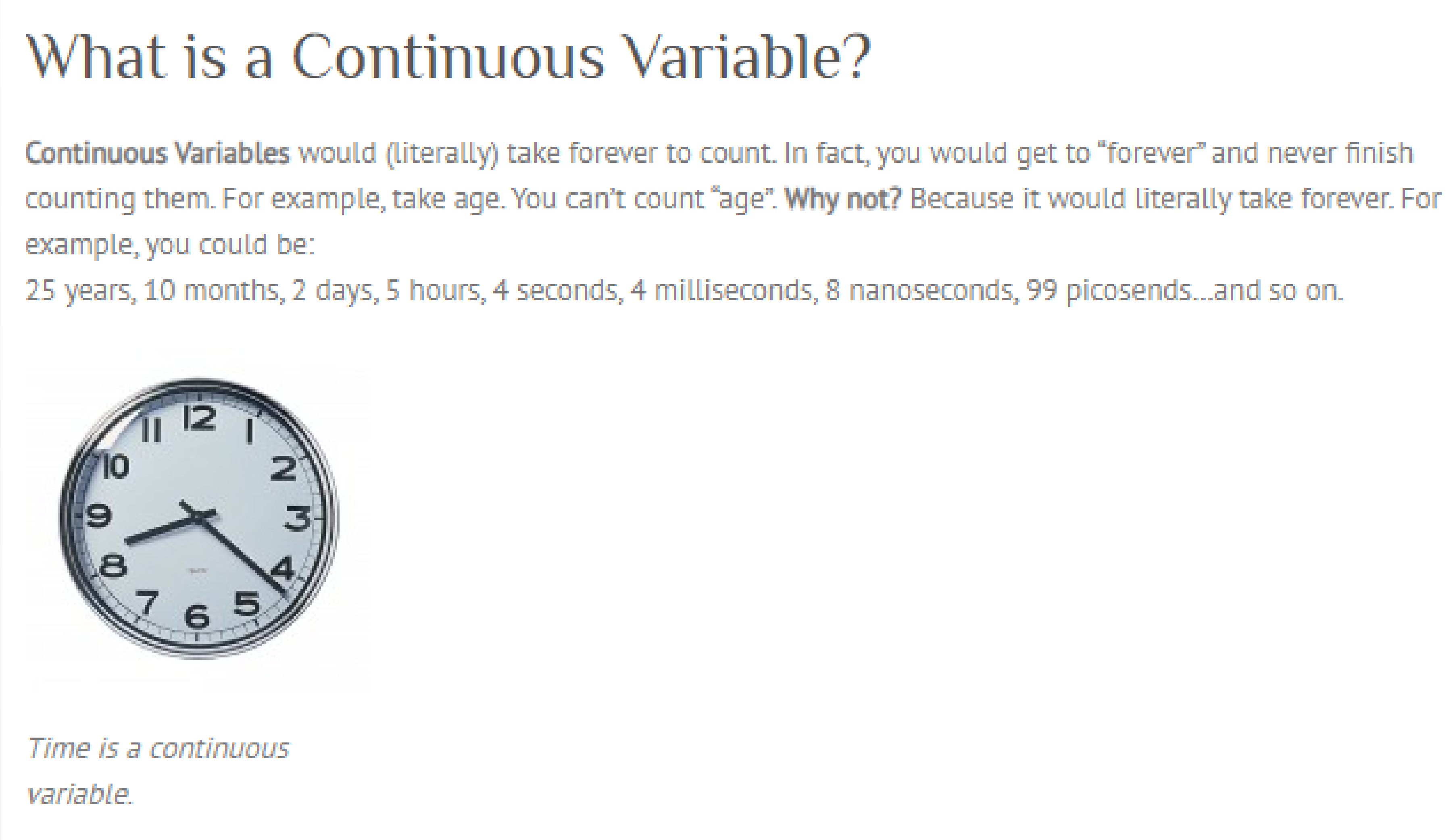
What is a Discrete Variable?

Discrete variables are countable in a finite amount of time. For example, you can count the change in your pocket. You can count the money in your bank account. You could also count the amount of money in *everyone's* bank accounts. It might take you a long time to count that last item, but the point is—it's still countable.

Discrete variables on a scatter plot.

quantitative continuous variables

- **quantitative continuous** variables are variables for which the values are **not countable** and have an **infinite number of possibilities**.
- For example:
  - Age Weight Height
- we usually referred to years, kilograms and centimeters for age, weight and height respectively.
- However, a 28-year-old man could actually be 28 years, 7 months, 16 days, 3 hours, 4 minutes, 5 seconds, 31 milliseconds, 9 nanoseconds old.
- For all measurements, we usually stop at a standard level of granularity, but nothing (except our measurement tools) prevents us from going deeper, leading to an **infinite number of potential values**.
- The fact that the values can take an infinite number of possibilities makes it uncountable.

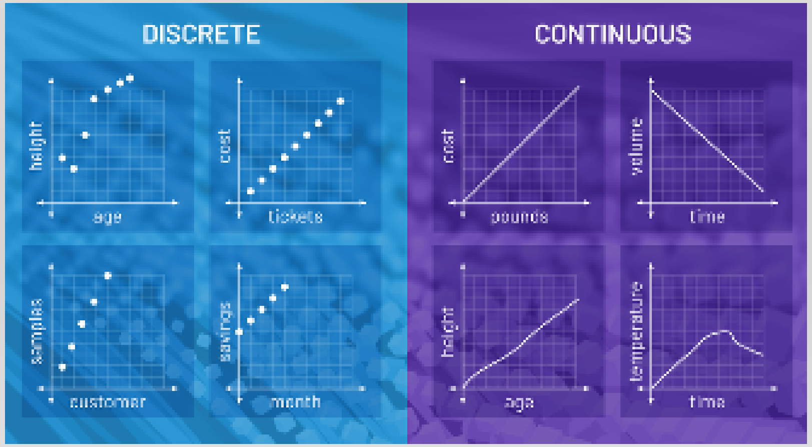


What is a Continuous Variable?

Continuous Variables would (literally) take forever to count. In fact, you would get to "forever" and never finish counting them. For example, take age. You can't count "age". **Why not?** Because it would literally take forever. For example, you could be:
25 years, 10 months, 2 days, 5 hours, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picosends…and so on.

Time is a continuous variable.

- A **discrete variable** is a variable whose value is obtained by counting.
    - number of students present
    - number of red marbles in a jar
    - number of heads when flipping three coins
    - students’ grade level
- A **continuous variable** is a variable whose value is obtained by measuring.
    - height of students in class
    - weight of students in class
    - time it takes to get to school
    - distance traveled between classes



Qualitative variables

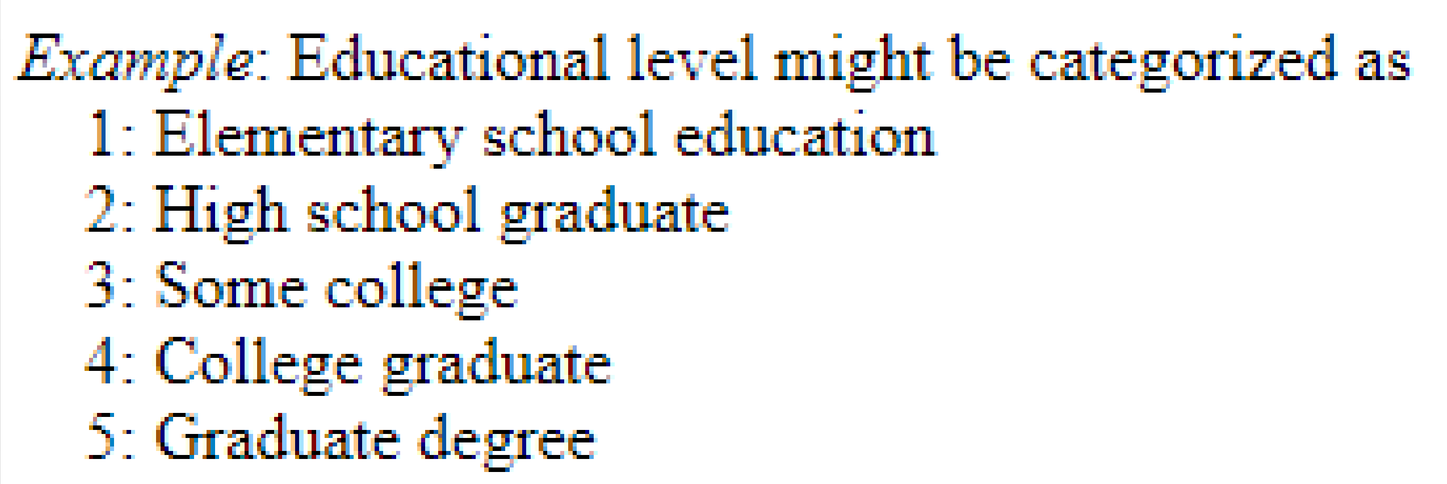
- **Qualitative** variables are variables that are **not numerical** and which **values fits into categories**.
- In other words, a **qualitative** variable is a variable which takes as its values modalities, **categories** or even levels, in contrast to **quantitative** variables which measure a **quantity** on each individual.
- Qualitative variables are divided into two types:
  - **nominal**
  - **ordinal**.

Qualitative nominal variable

- A **qualitative nominal** variable is a qualitative variable where **no ordering** is possible or implied in the levels.
- For example, the variable gender is nominal because there is no order in the levels female/male.
- Eye color is another example of a nominal variable because there is no order among blue, brown or green eyes.
- A nominal variable can have between two levels
  - (what is your gender? Female/Male)
- and a large number of levels
  - (what is your college major? Each major is a level in that case).

Qualitative ordinal variable

- **Qualitative ordinal** variable is a qualitative variable with an **order implied in the levels**.
- good example is health, which can take values such as poor, reasonable, good, or excellent.



Example: Educational level might be categorized as

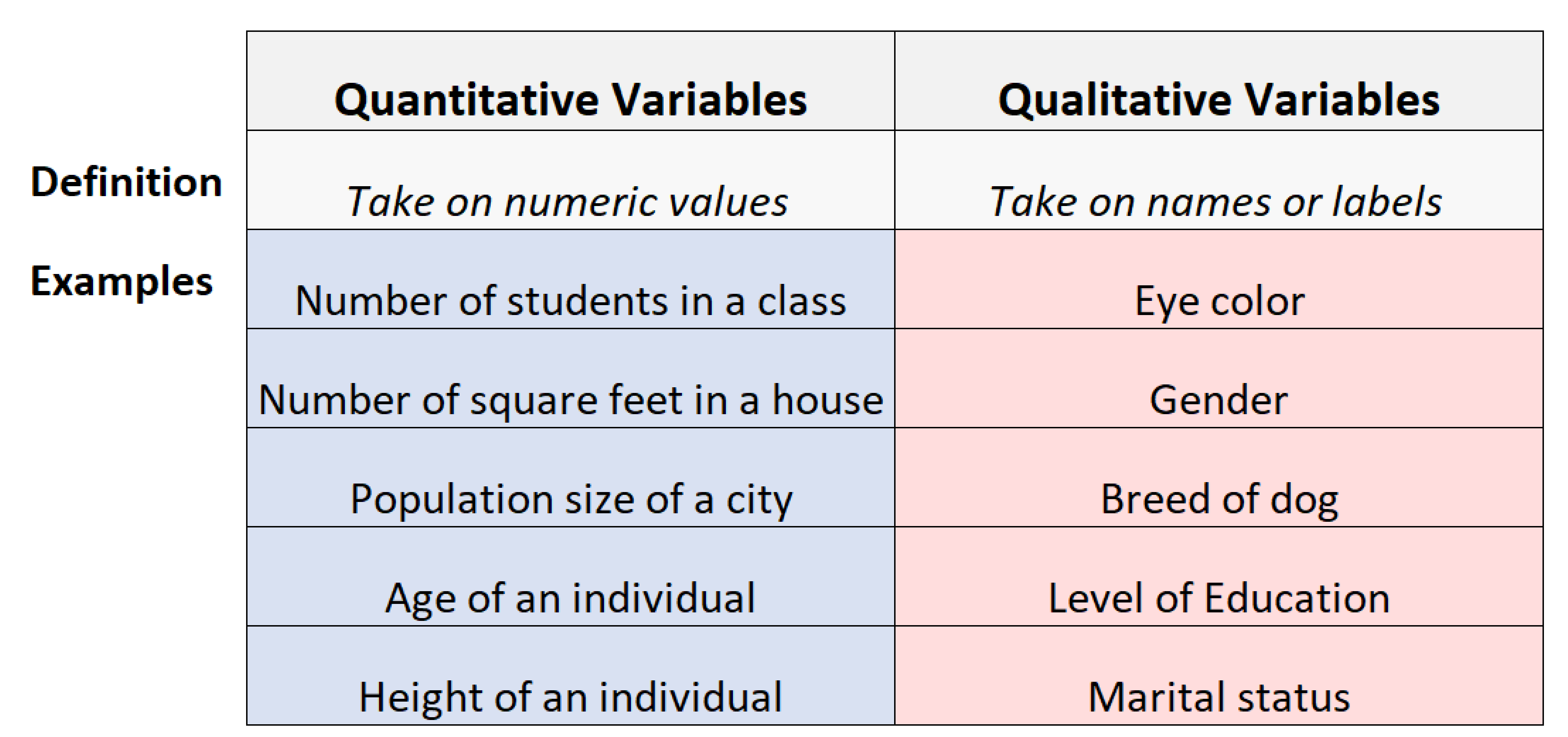
1: Elementary school education

2: High school graduate

3: Some college

4: College graduate

5: Graduate degree



	Quantitative Variables	**Qualitative Variables**
Definition	*Take on numeric values*	*Take on names or labels*
Examples	Number of students in a class	Eye color
	Number of square feet in a house	Gender
	Population size of a city	Breed of dog
	Age of an individual	Level of Education
	Height of an individual	Marital status

LEVELS OF MEASUREMENT

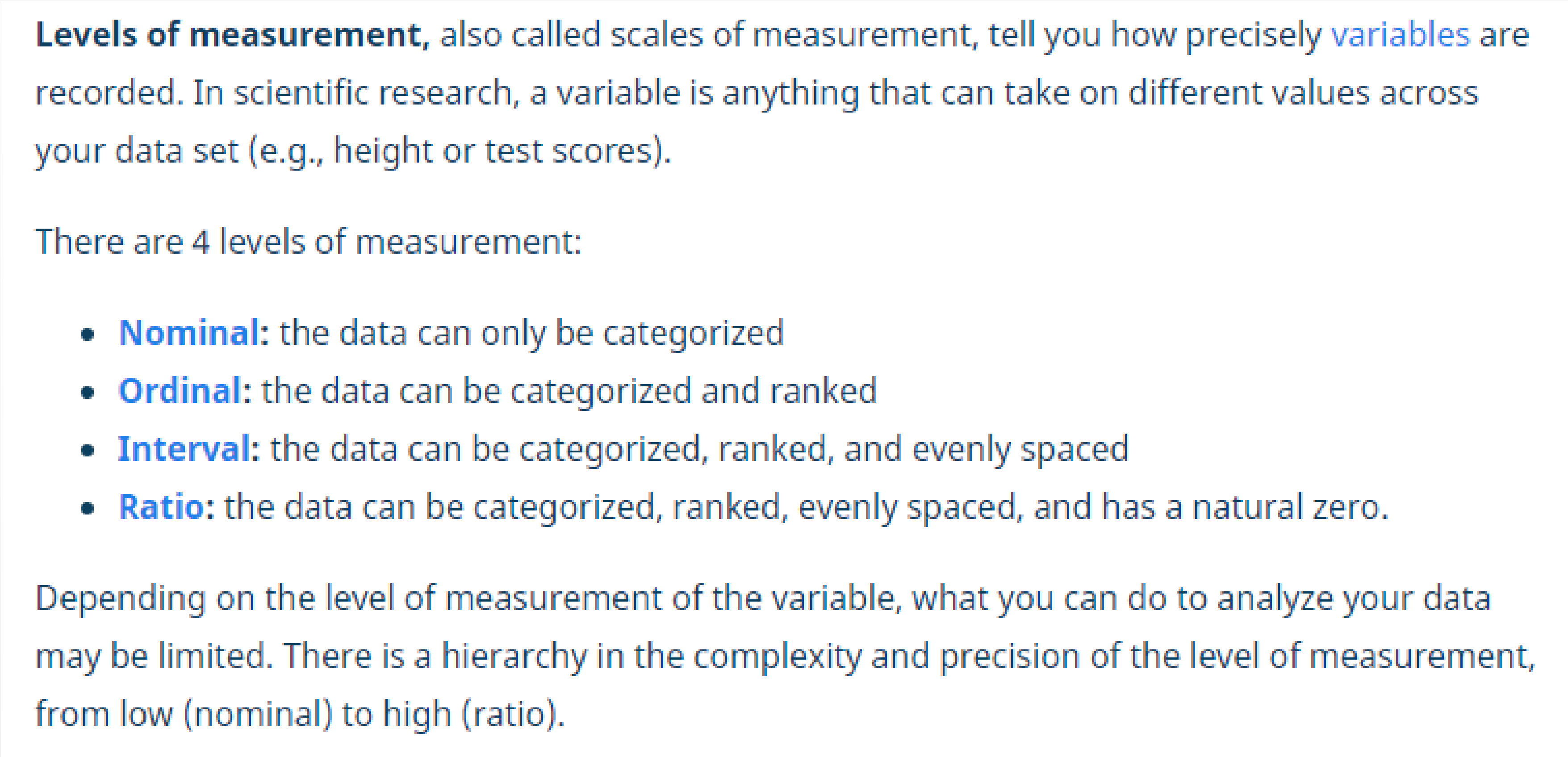
- When we talk about levels of measurement, we are talking about how we measure a variable.
- Variables have four different levels of measurement:
  - **Nominal**
  - **Ordinal**
  - **Interval**
  - **Ratio**

- **Nominal** variables are *categorical* variables where the categories are different only because they are named differently.
    - We cannot rank or order the categories.
    - Some examples include the following: race/ethnicity, gender, eye color, or neighborhood.

- **Ordinal** variables are *categorical* variables where the categories can be ordered or ranked.
    - Some examples include the following: class level (freshman, sophomore, junior, senior) and education level (less than HS, HS diploma, some college, college degree).

- **Interval** variables are *continuous/scale* variables with no meaningful/absolute zero.
  - A meaningful/absolute zero means that there is an absence of something.
  - In an interval variable, 0 is just another data point along the scale, it does NOT mean the absence of something.
  - For example, 0 degrees Fahrenheit is not the absence of heat or temperature, it is just another number along the temperature spectrum (it does mean it’s pretty cold, though).

- **Ratio** variables are *continuous/scale* variables with a meaningful/absolute zero.
  - In a ratio variable, 0 means that there is nothing there.
  - For example, if I have 0 dollars, I have no money. If I have 0 hairs on my head, I am bald.

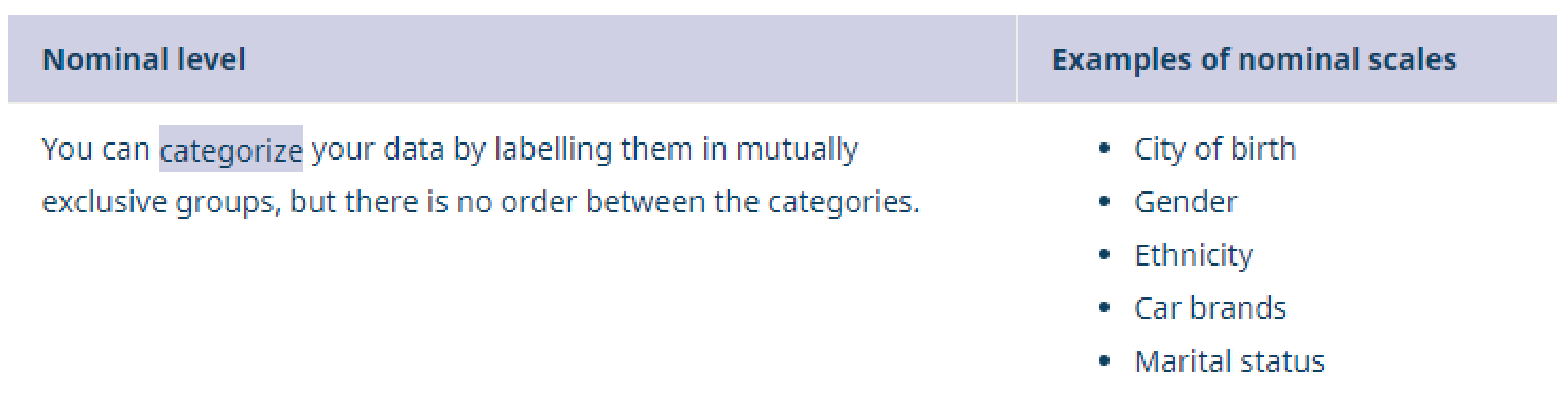


Levels of measurement, also called scales of measurement, tell you how precisely variables are recorded. In scientific research, a variable is anything that can take on different values across your data set (e.g., height or test scores).

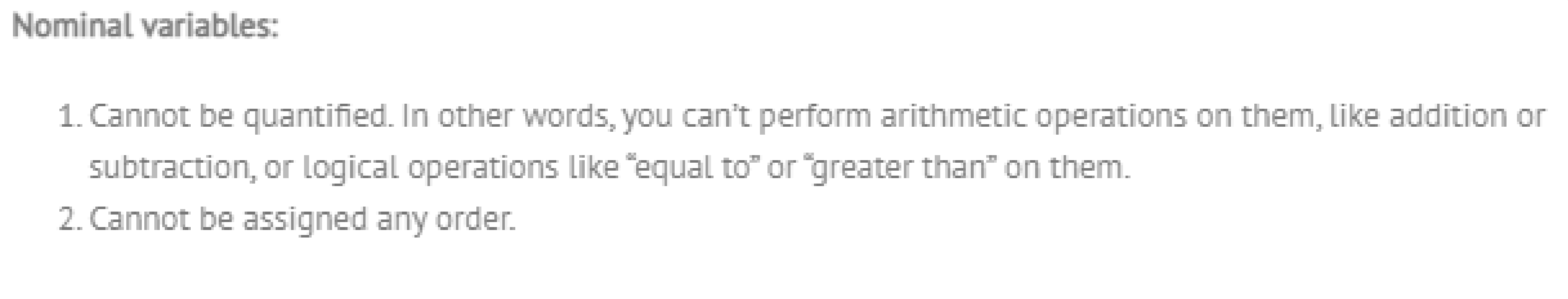
There are 4 levels of measurement:

- **Nominal:** the data can only be categorized
- **Ordinal:** the data can be categorized and ranked
- **Interval:** the data can be categorized, ranked, and evenly spaced
- **Ratio:** the data can be categorized, ranked, evenly spaced, and has a natural zero.

Depending on the level of measurement of the variable, what you can do to analyze your data may be limited. There is a hierarchy in the complexity and precision of the level of measurement, from low (nominal) to high (ratio).

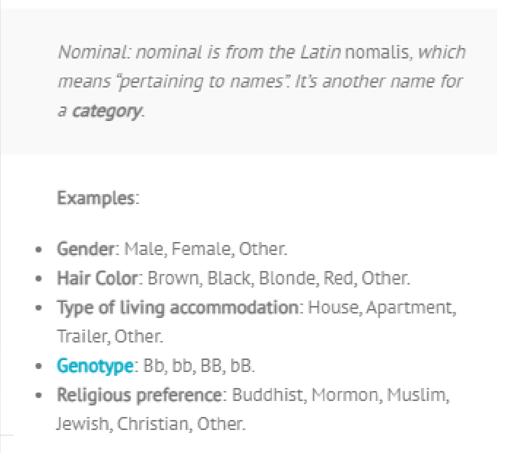


Nominal level	Examples of nominal scales
You can categorize your data by labelling them in mutually exclusive groups, but there is no order between the categories.	• City of birth • Gender • Ethnicity • Car brands • Marital status



Nominal variables:

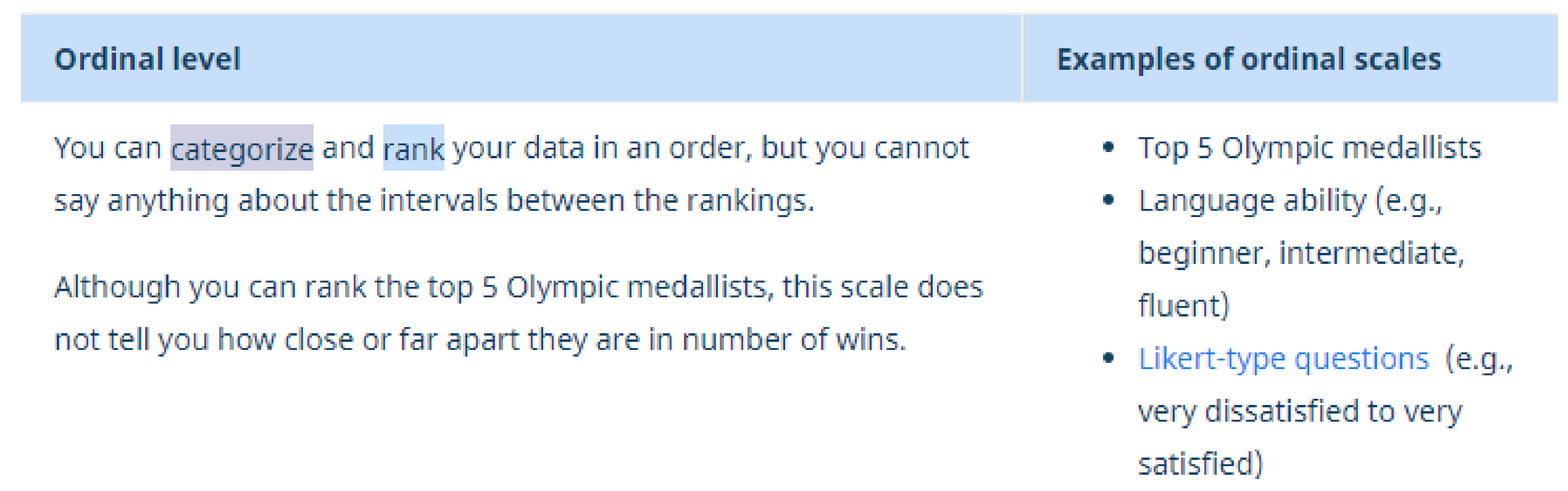
1. Cannot be quantified. In other words, you can't perform arithmetic operations on them, like addition or subtraction, or logical operations like "equal to" or "greater than" on them.
2. Cannot be assigned any order.



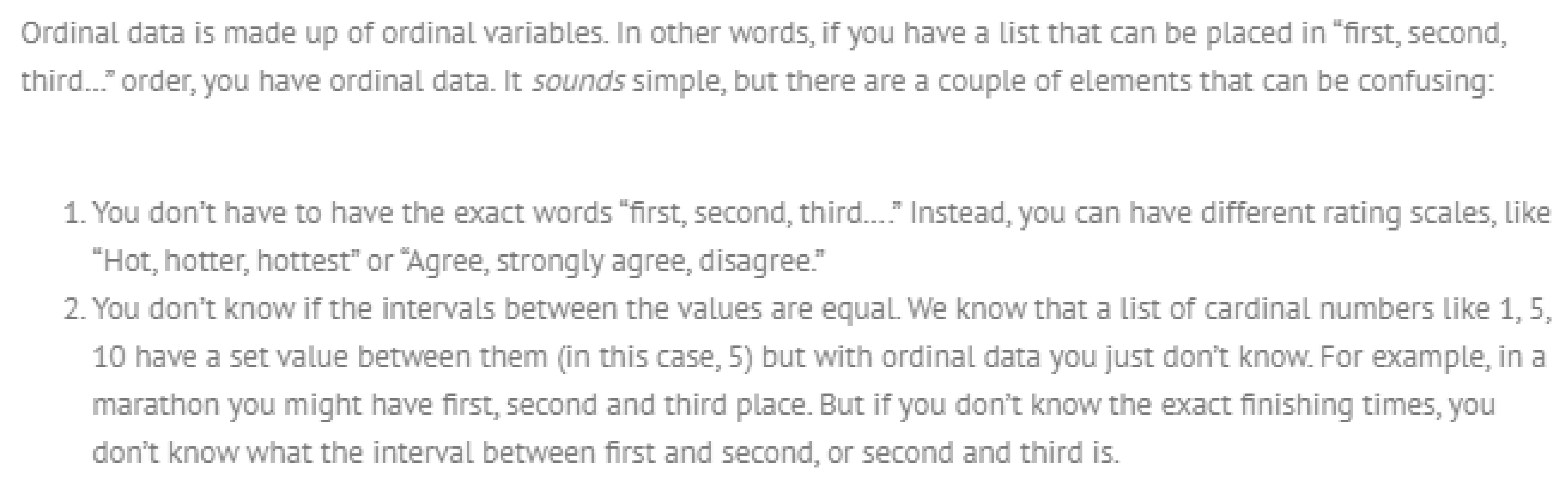
Nominal: nominal is from the Latin nomalis, *which means "pertaining to names". It's another name for a* ***category****.*

Examples:

- **Gender**: Male, Female, Other.
- **Hair Color**: Brown, Black, Blonde, Red, Other.
- **Type of living accommodation**: House, Apartment, Trailer, Other.
- **Genotype**: Bb, bb, BB, bB.
- **Religious preference**: Buddhist, Mormon, Muslim, Jewish, Christian, Other.

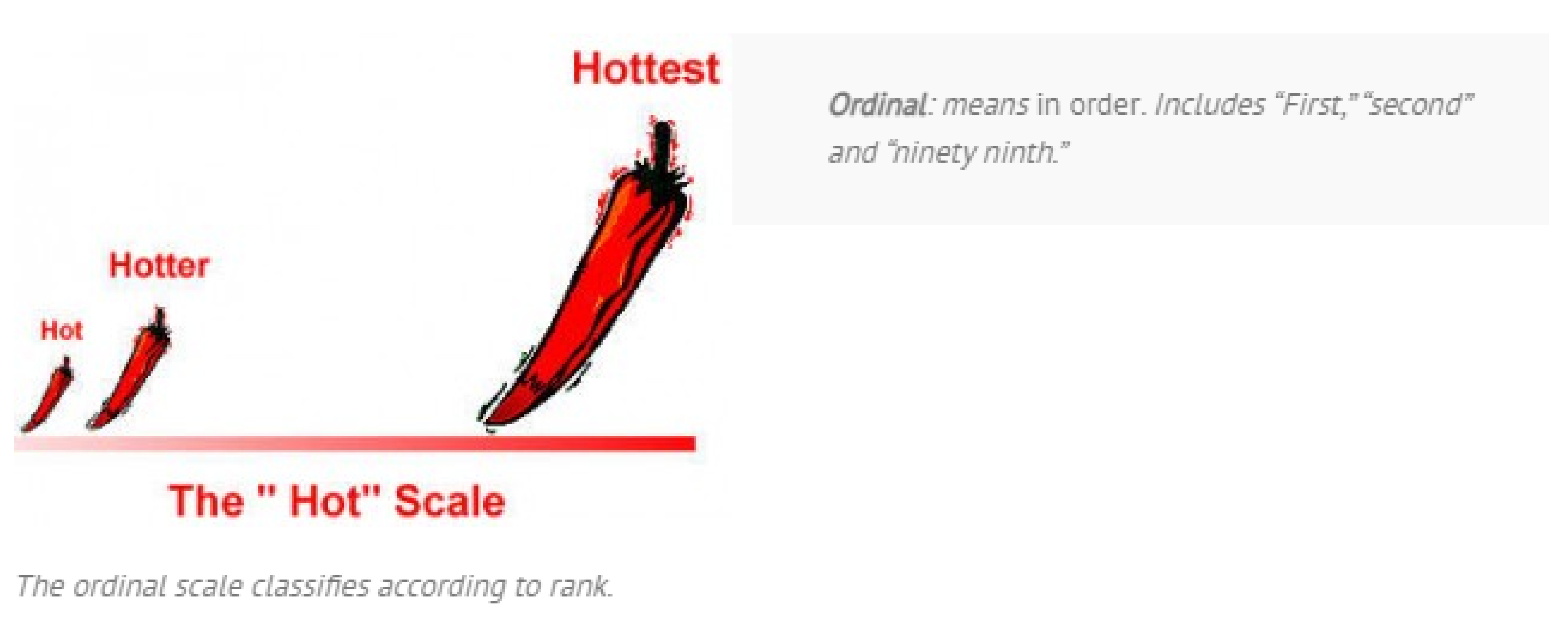


Ordinal level	Examples of ordinal scales
You can categorize and rank your data in an order, but you cannot say anything about the intervals between the rankings. Although you can rank the top 5 Olympic medallists, this scale does not tell you how close or far apart they are in number of wins.	• Top 5 Olympic medallists • Language ability (e.g., beginner, intermediate, fluent) • Likert-type questions (e.g., very dissatisfied to very satisfied)



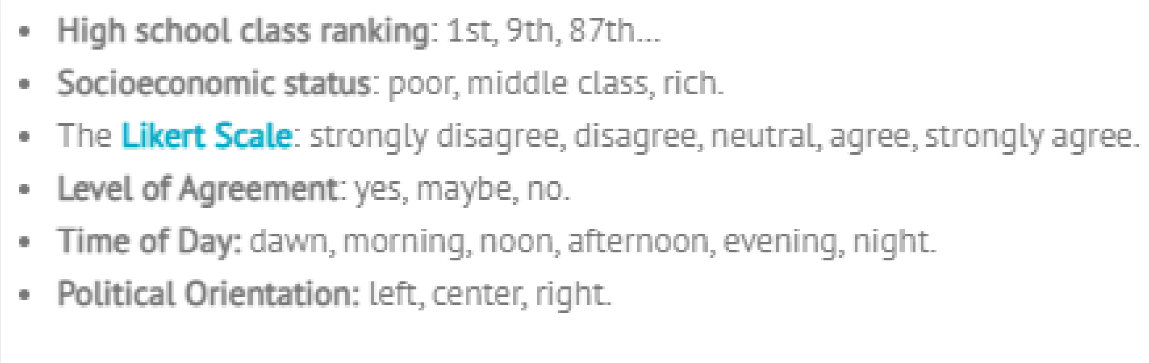
Ordinal data is made up of ordinal variables. In other words, if you have a list that can be placed in "first, second, third..." order, you have ordinal data. It *sounds* simple, but there are a couple of elements that can be confusing:

1. You don't have to have the exact words "first, second, third...." Instead, you can have different rating scales, like "Hot, hotter, hottest" or "Agree, strongly agree, disagree."
2. You don't know if the intervals between the values are equal. We know that a list of cardinal numbers like 1, 5, 10 have a set value between them (in this case, 5) but with ordinal data you just don't know. For example, in a marathon you might have first, second and third place. But if you don't know the exact finishing times, you don't know what the interval between first and second, or second and third is.

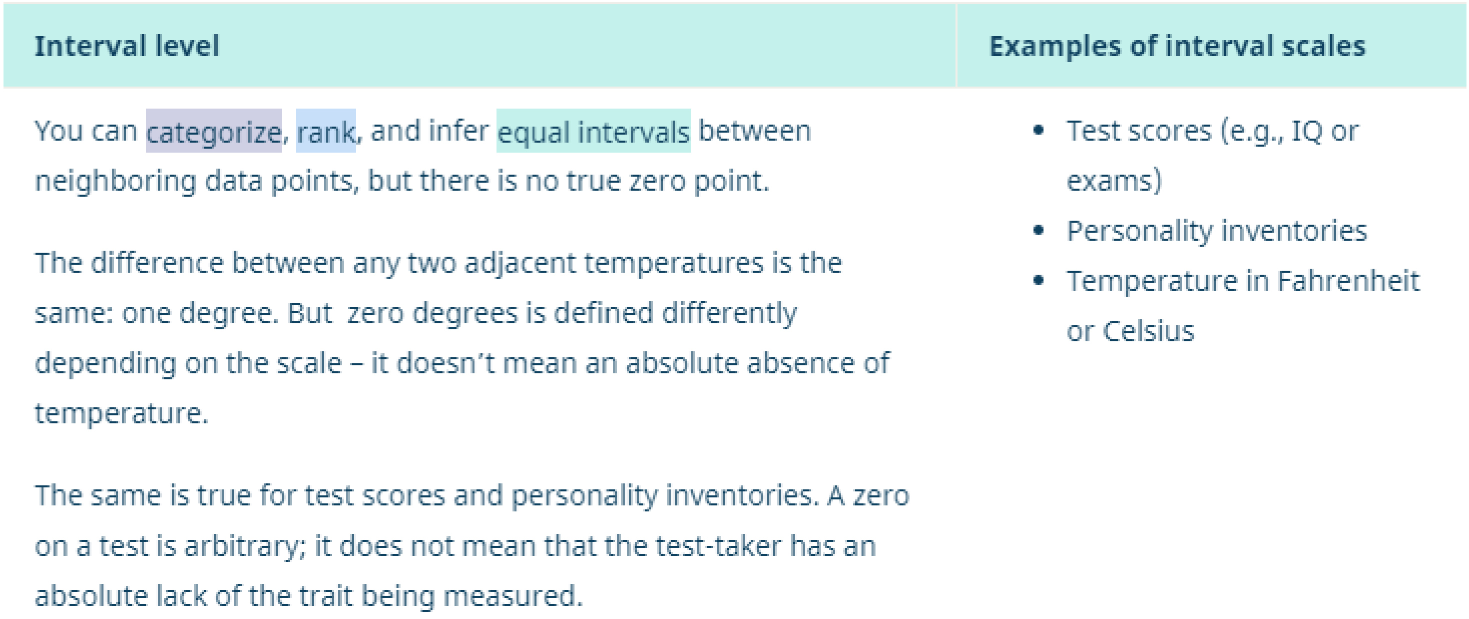


***Ordinal**: means* in order. *Includes "First," "second" and "ninety ninth."*

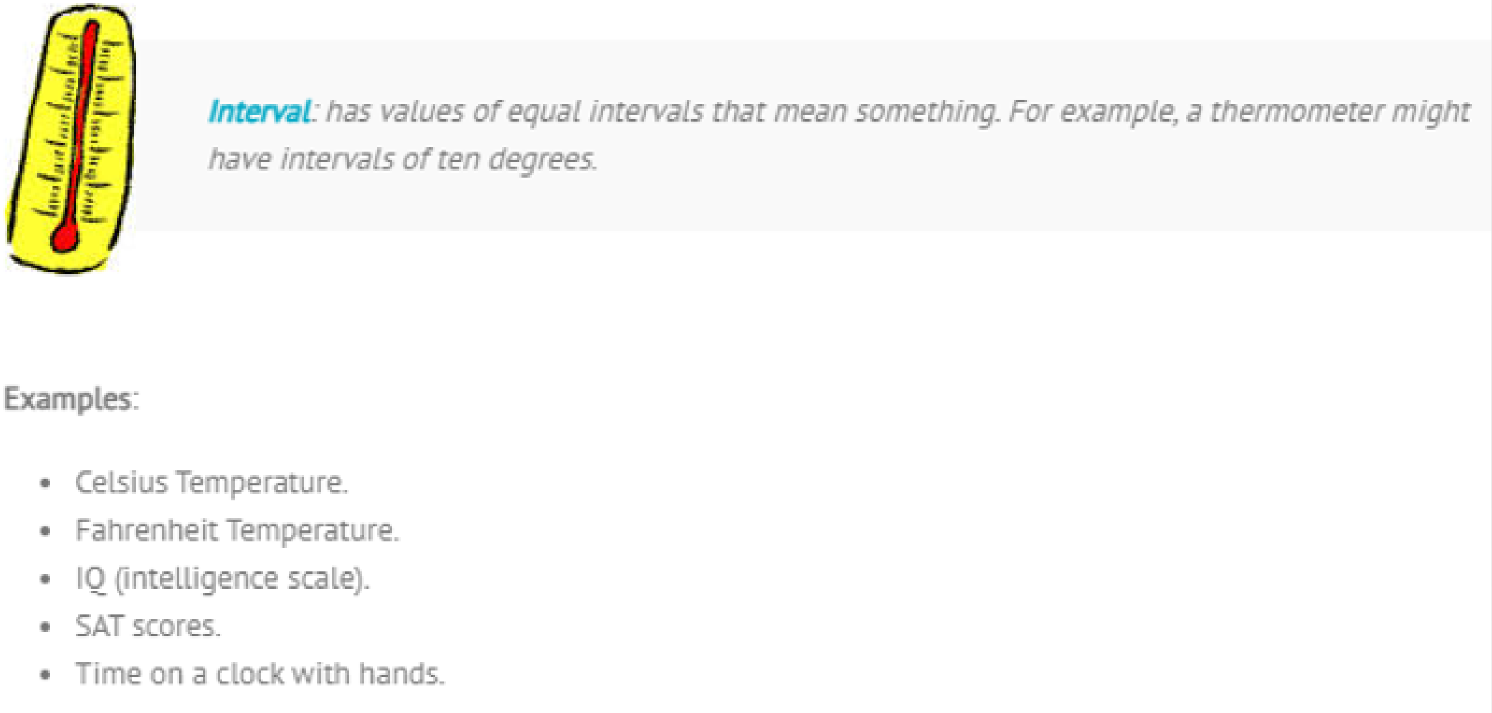
The ordinal scale classifies according to rank.



- **High school class ranking**: 1st, 9th, 87th…
- **Socioeconomic status**: poor, middle class, rich.
- The **Likert Scale**: strongly disagree, disagree, neutral, agree, strongly agree.
- **Level of Agreement**: yes, maybe, no.
- **Time of Day:** dawn, morning, noon, afternoon, evening, night.
- **Political Orientation:** left, center, right.



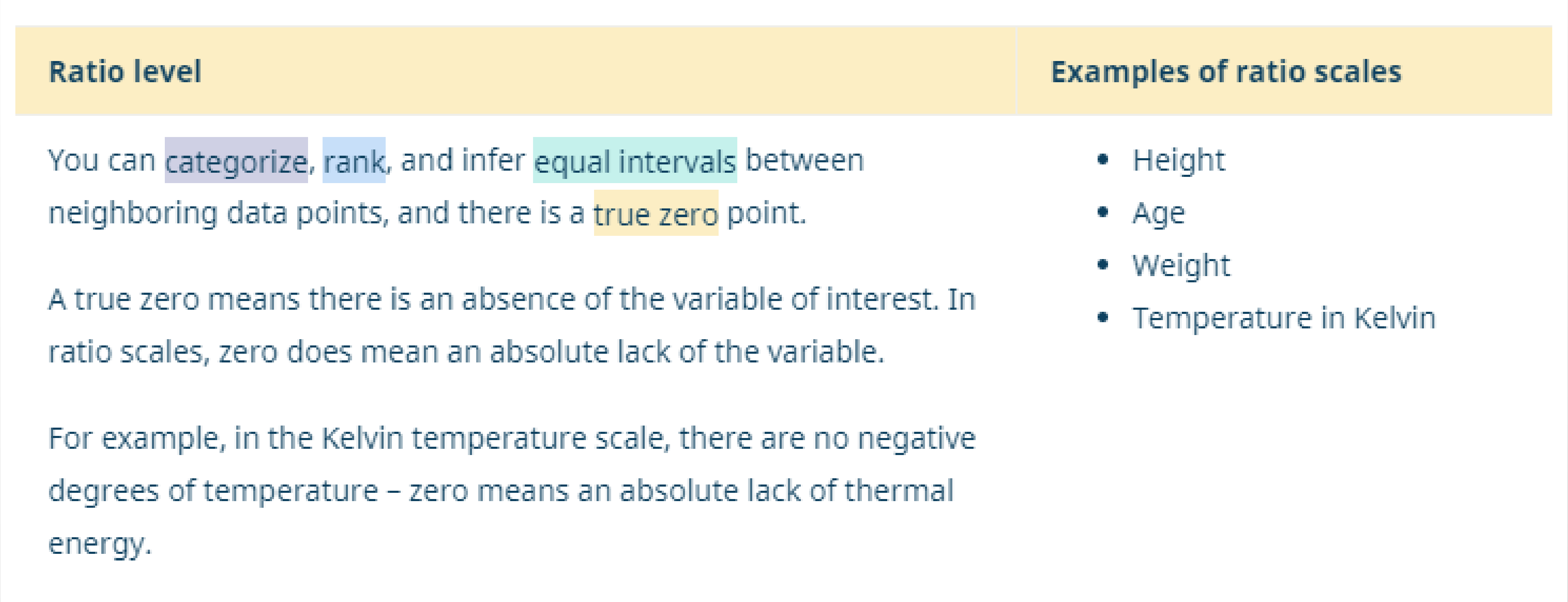
Interval level	Examples of interval scales
You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point. The difference between any two adjacent temperatures is the same: one degree. But zero degrees is defined differently depending on the scale – it doesn’t mean an absolute absence of temperature. The same is true for test scores and personality inventories. A zero on a test is arbitrary; it does not mean that the test-taker has an absolute lack of the trait being measured.	• Test scores (e.g., IQ or exams) • Personality inventories • Temperature in Fahrenheit or Celsius



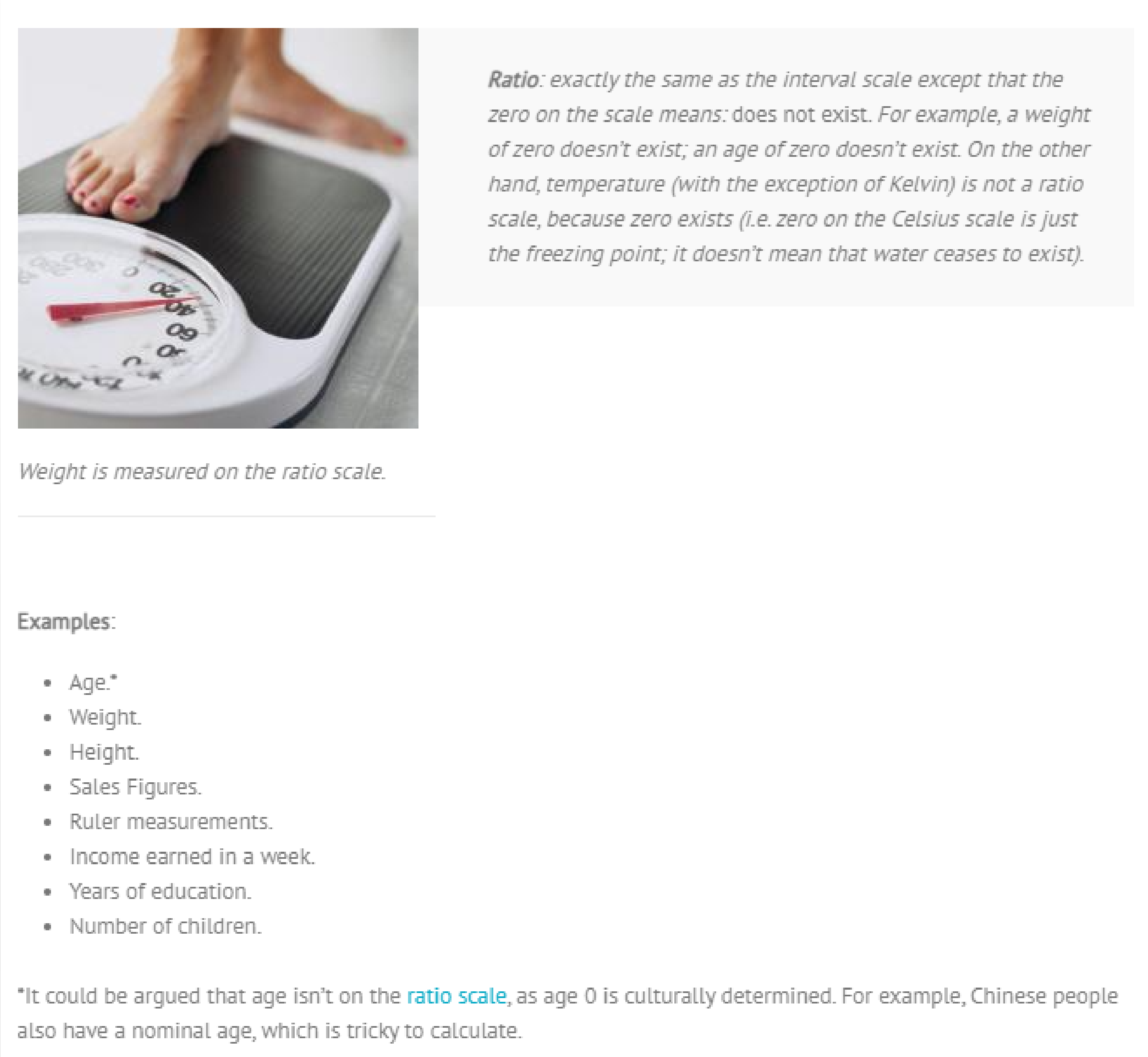
Interval: has values of equal intervals that mean something. For example, a thermometer might have intervals of ten degrees.

Examples:

- Celsius Temperature.
- Fahrenheit Temperature.
- IQ (intelligence scale).
- SAT scores.
- Time on a clock with hands.



Ratio level	Examples of ratio scales
You can categorize, rank, and infer equal intervals between neighboring data points, and there is a true zero point. A true zero means there is an absence of the variable of interest. In ratio scales, zero does mean an absolute lack of the variable. For example, in the Kelvin temperature scale, there are no negative degrees of temperature – zero means an absolute lack of thermal energy.	• Height • Age • Weight • Temperature in Kelvin



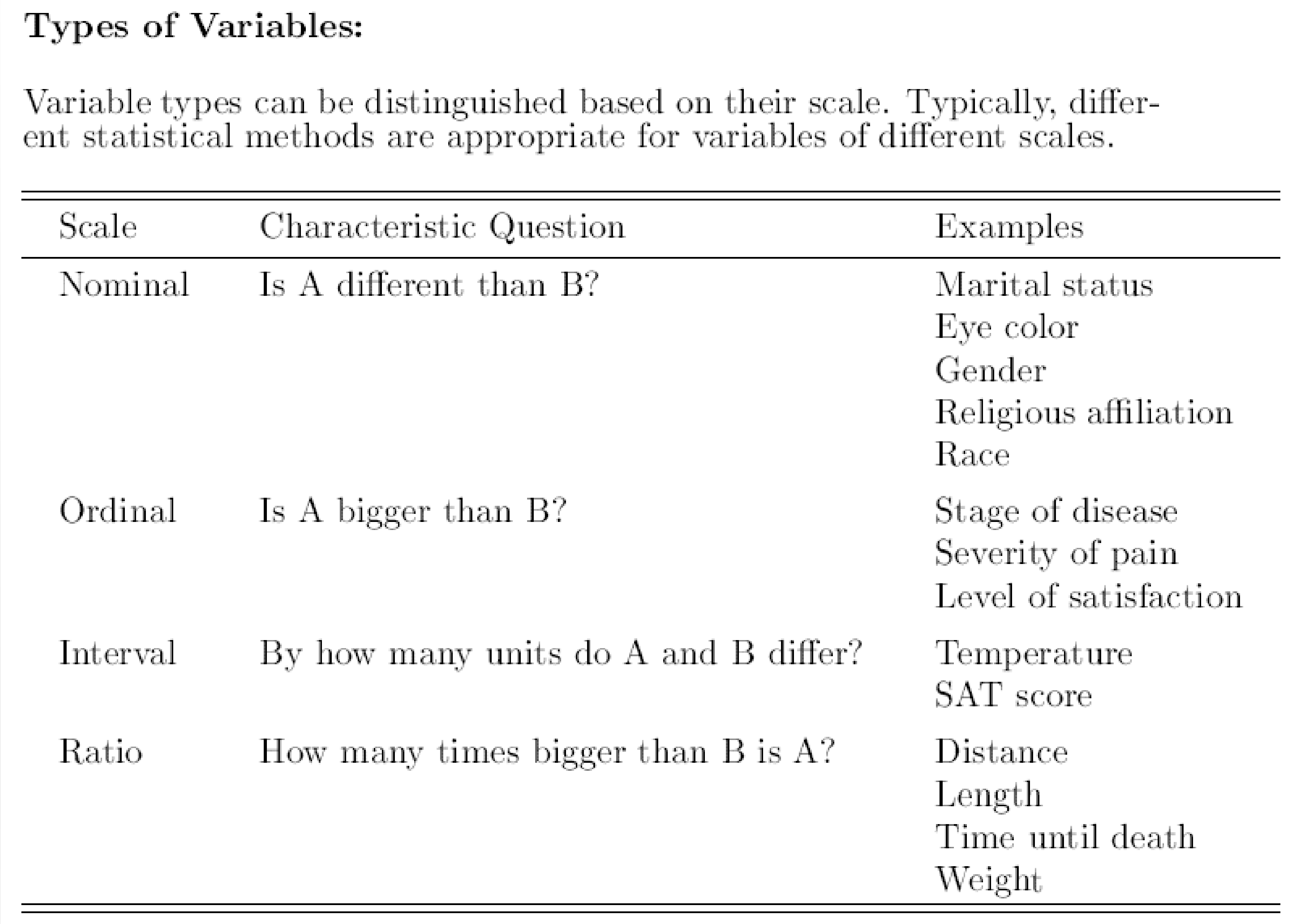
Ratio*: exactly the same as the interval scale except that the zero on the scale means:* does not exist. *For example, a weight of zero doesn't exist; an age of zero doesn't exist. On the other hand, temperature (with the exception of Kelvin) is not a ratio scale, because zero exists (i.e. zero on the Celsius scale is just the freezing point; it doesn't mean that water ceases to exist).*

Weight is measured on the ratio scale.

Examples:

- Age.*
- Weight.
- Height.
- Sales Figures.
- Ruler measurements.
- Income earned in a week.
- Years of education.
- Number of children.

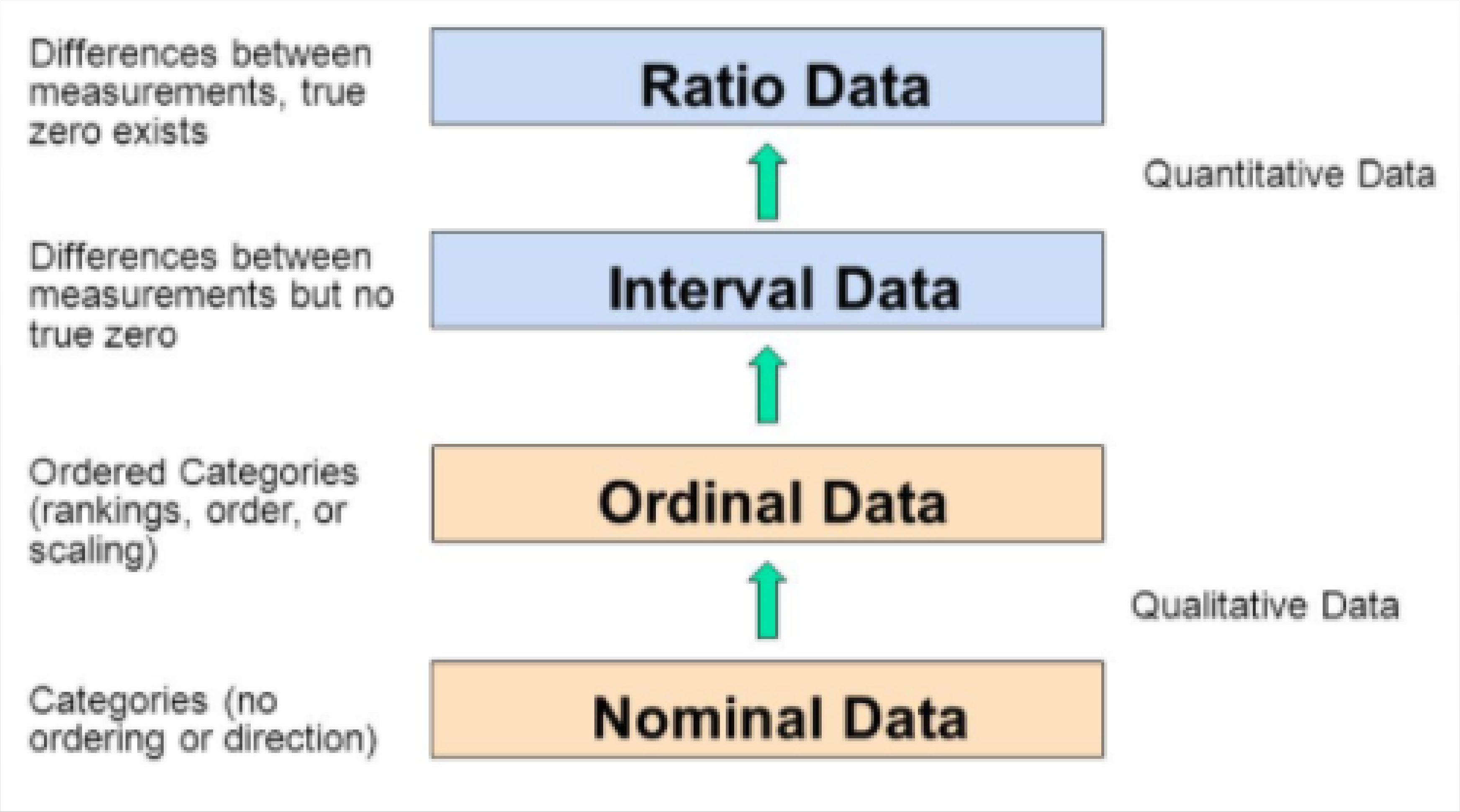
*It could be argued that age isn't on the ratio scale, as age 0 is culturally determined. For example, Chinese people also have a nominal age, which is tricky to calculate.



Types of Variables:

Variable types can be distinguished based on their scale. Typically, different statistical methods are appropriate for variables of different scales.

Scale	Characteristic Question	Examples
Nominal	Is A different than B?	Marital status Eye color Gender Religious affiliation Race
Ordinal	Is A bigger than B?	Stage of disease Severity of pain Level of satisfaction
Interval	By how many units do A and B differ?	Temperature SAT score
Ratio	How many times bigger than B is A?	Distance Length Time until death Weight



Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

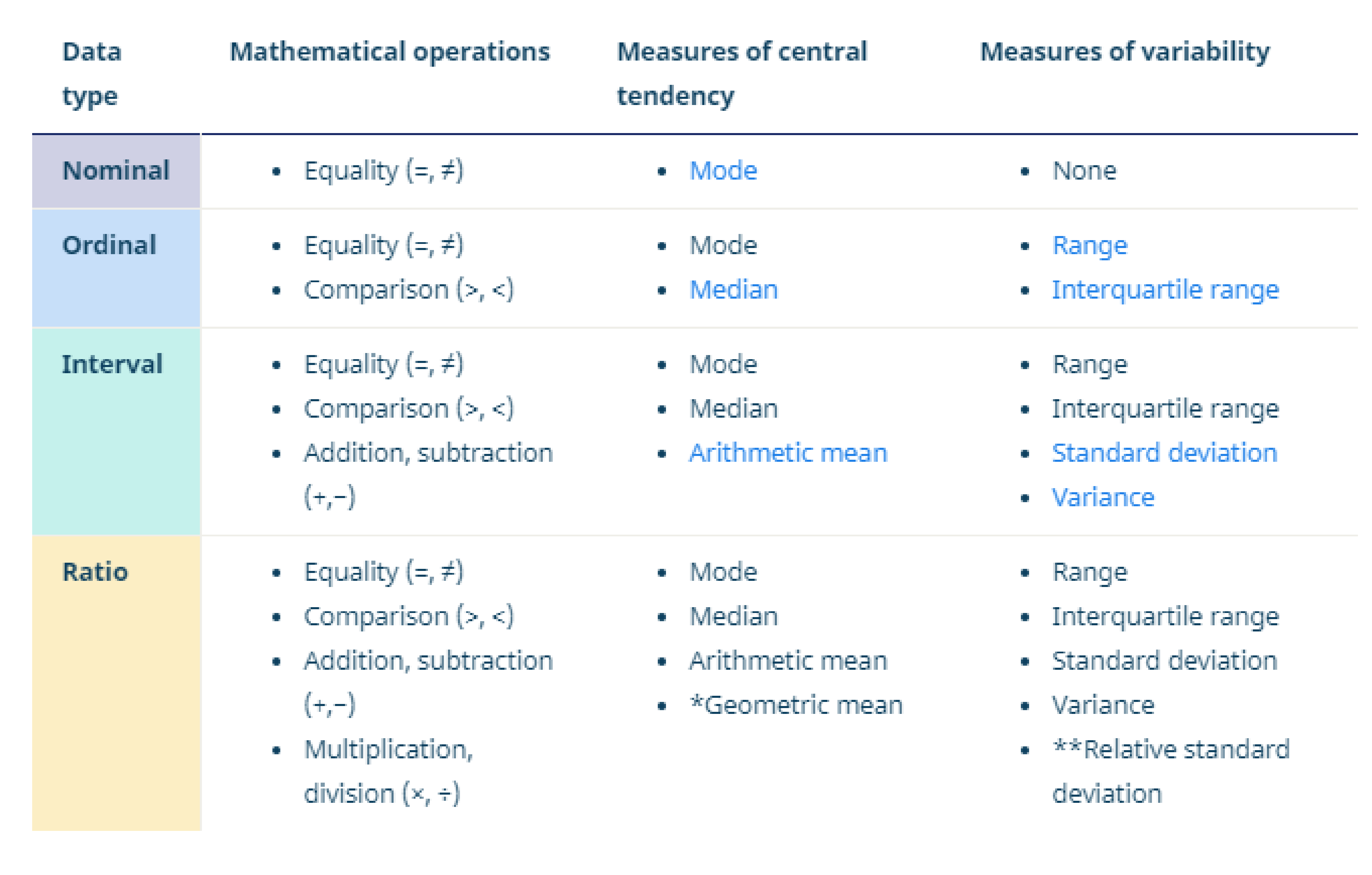
Ordered Categories (rankings, order, or scaling)

Ordinal Data

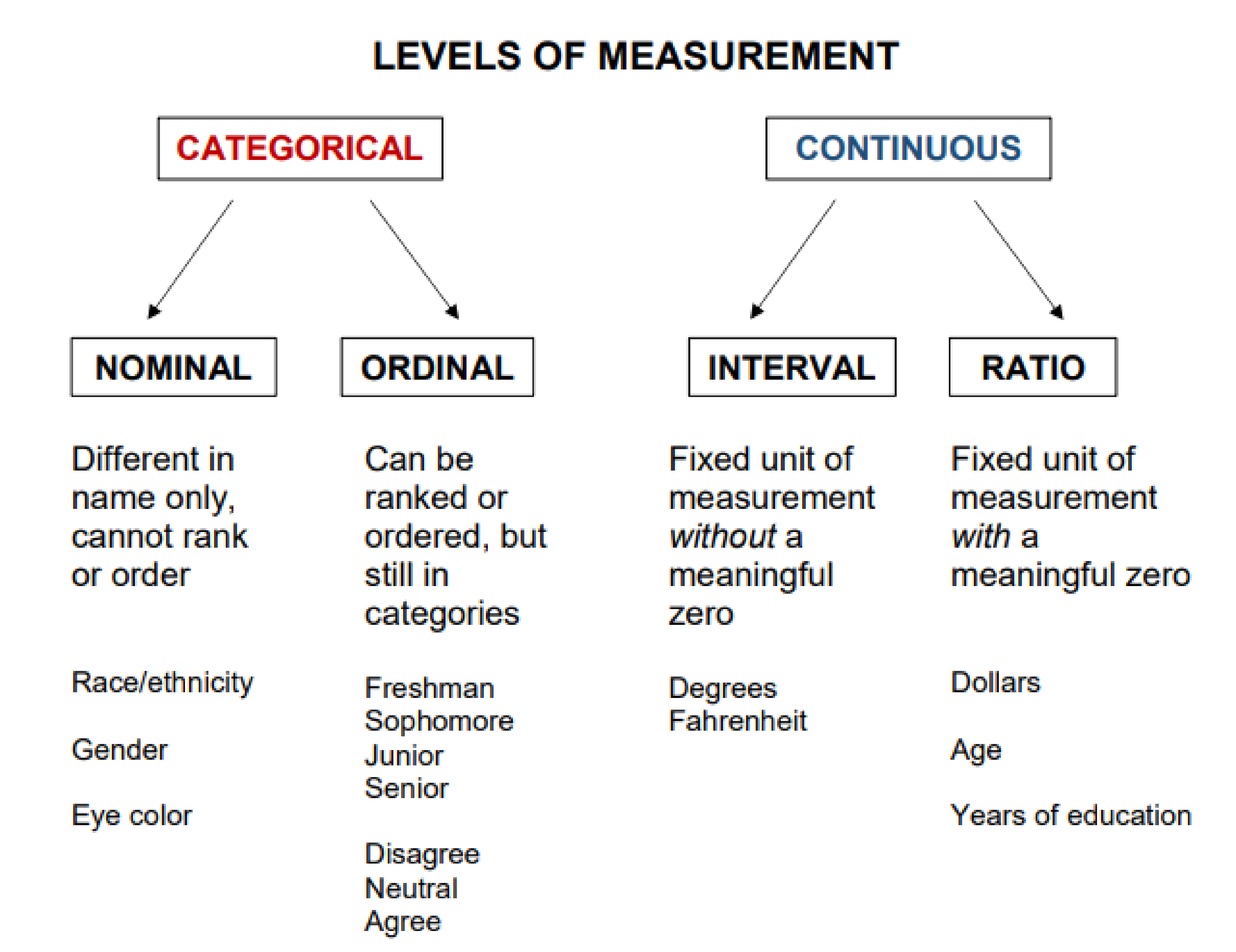
Qualitative Data

Categories (no ordering or direction)

Nominal Data



Data type	Mathematical operations	Measures of central tendency	Measures of variability
Nominal	• Equality (=, ≠)	• Mode	• None
Ordinal	• Equality (=, ≠) • Comparison (>, <)	• Mode • Median	• Range • Interquartile range
Interval	• Equality (=, ≠) • Comparison (>, <) • Addition, subtraction (+,−)	• Mode • Median • Arithmetic mean	• Range • Interquartile range • Standard deviation • Variance
Ratio	• Equality (=, ≠) • Comparison (>, <) • Addition, subtraction (+,−) • Multiplication, division (×, ÷)	• Mode • Median • Arithmetic mean • *Geometric mean	• Range • Interquartile range • Standard deviation • Variance • **Relative standard deviation



LEVELS OF MEASUREMENT

CATEGORICAL

NOMINAL

Different in name only, cannot rank or order

Race/ethnicity

Gender

Eye color

ORDINAL

Can be ranked or ordered, but still in categories

Freshman
Sophomore
Junior
Senior

Disagree
Neutral
Agree

CONTINUOUS

INTERVAL

Fixed unit of measurement *without* a meaningful zero

Degrees Fahrenheit

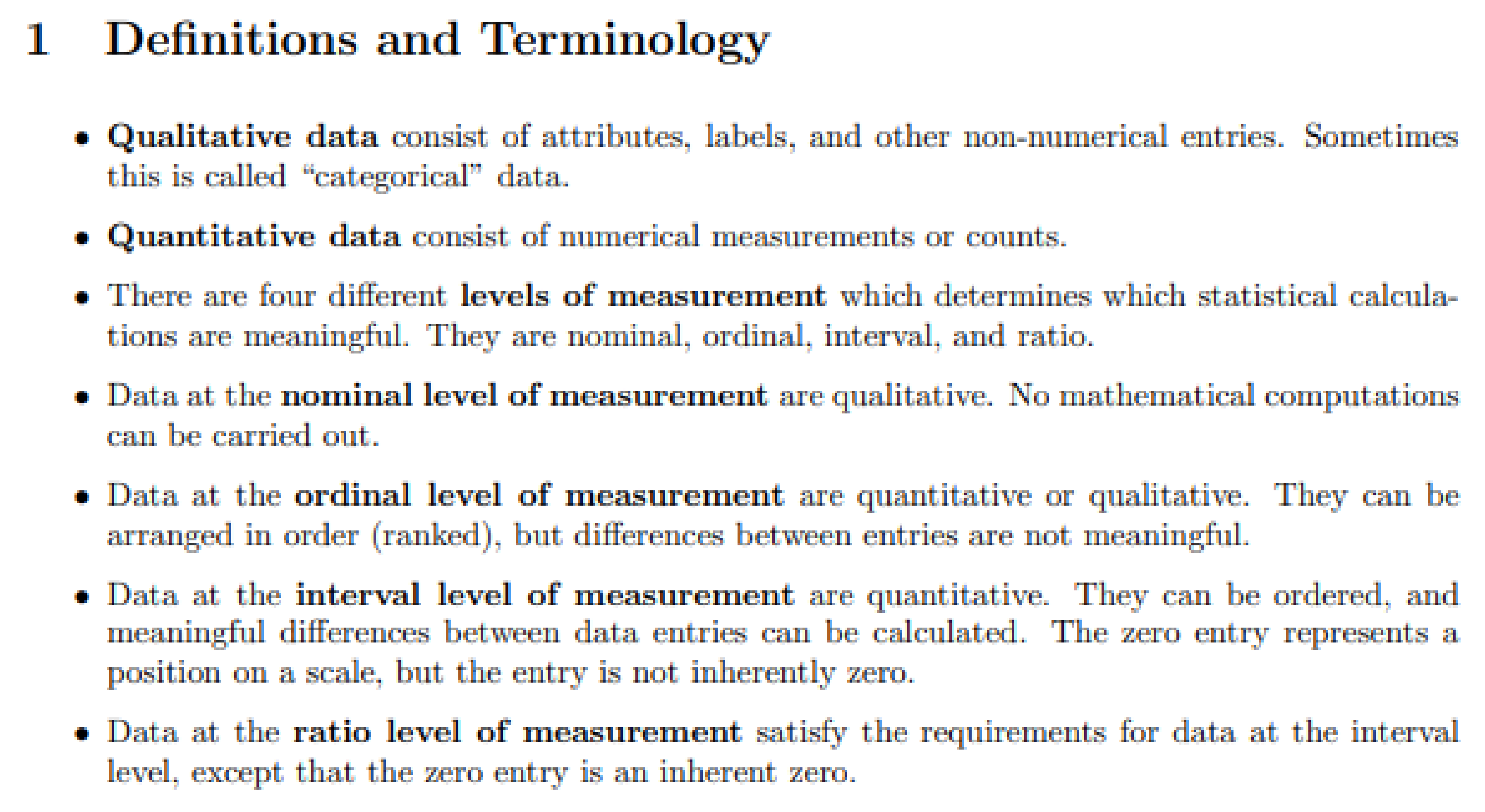
RATIO

Fixed unit of measurement *with* a meaningful zero

Dollars

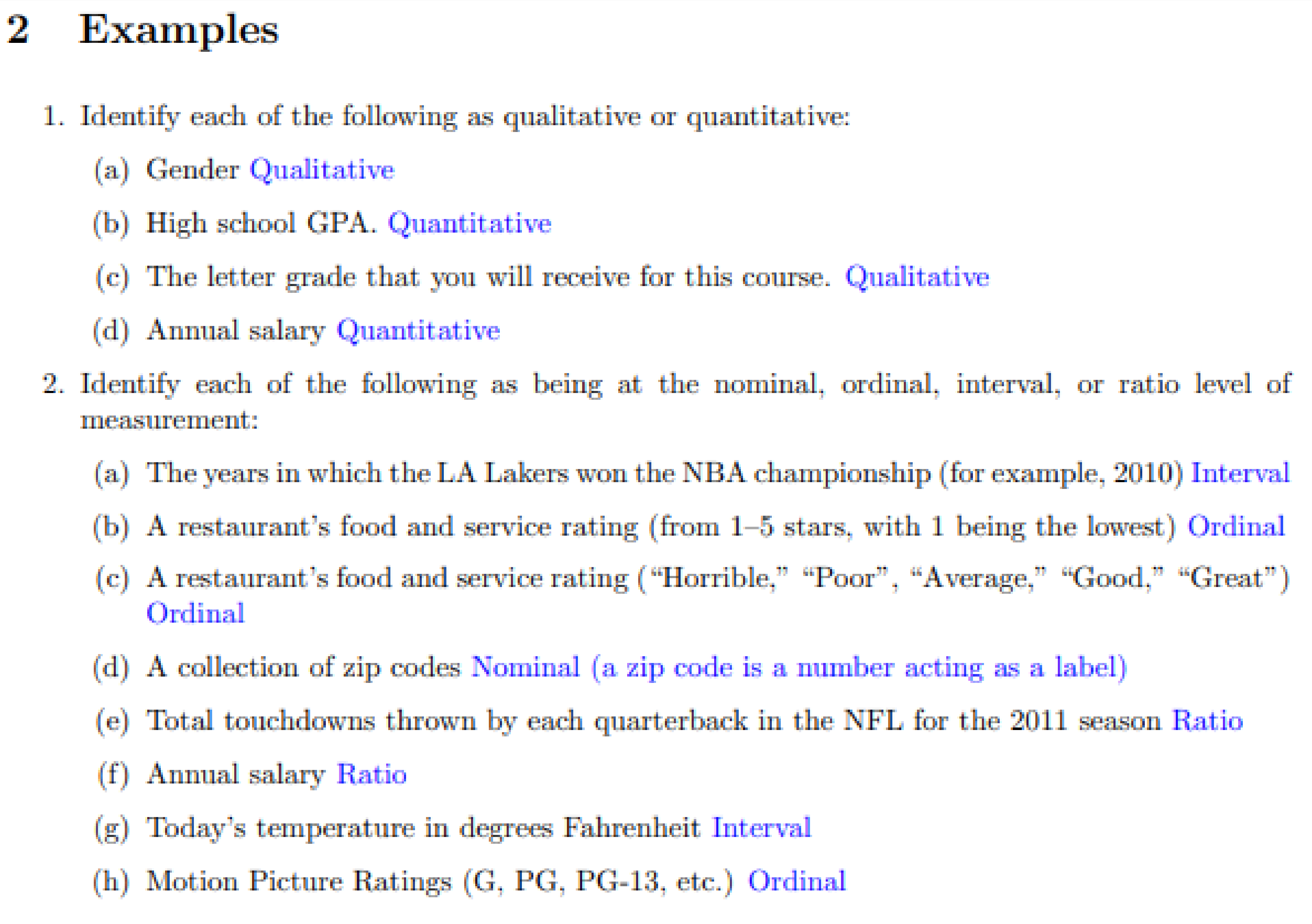
Age

Years of education



1 Definitions and Terminology

- **Qualitative data** consist of attributes, labels, and other non-numerical entries. Sometimes this is called "categorical" data.
- **Quantitative data** consist of numerical measurements or counts.
- There are four different **levels of measurement** which determines which statistical calculations are meaningful. They are nominal, ordinal, interval, and ratio.
- Data at the **nominal level of measurement** are qualitative. No mathematical computations can be carried out.
- Data at the **ordinal level of measurement** are quantitative or qualitative. They can be arranged in order (ranked), but differences between entries are not meaningful.
- Data at the **interval level of measurement** are quantitative. They can be ordered, and meaningful differences between data entries can be calculated. The zero entry represents a position on a scale, but the entry is not inherently zero.
- Data at the **ratio level of measurement** satisfy the requirements for data at the interval level, except that the zero entry is an inherent zero.



2 Examples

1. Identify each of the following as qualitative or quantitative:
   (a) Gender Qualitative
   (b) High school GPA. Quantitative
   (c) The letter grade that you will receive for this course. Qualitative
   (d) Annual salary Quantitative
2. Identify each of the following as being at the nominal, ordinal, interval, or ratio level of measurement:
   (a) The years in which the LA Lakers won the NBA championship (for example, 2010) Interval
   (b) A restaurant's food and service rating (from 1–5 stars, with 1 being the lowest) Ordinal
   (c) A restaurant's food and service rating ("Horrible," "Poor", "Average," "Good," "Great") Ordinal
   (d) A collection of zip codes Nominal (a zip code is a number acting as a label)
   (e) Total touchdowns thrown by each quarterback in the NFL for the 2011 season Ratio
   (f) Annual salary Ratio
   (g) Today's temperature in degrees Fahrenheit Interval
   (h) Motion Picture Ratings (G, PG, PG-13, etc.) Ordinal

QUIZ I. Determine the level of measurement. (Nominal, Ordinal, Interval, Ratio)

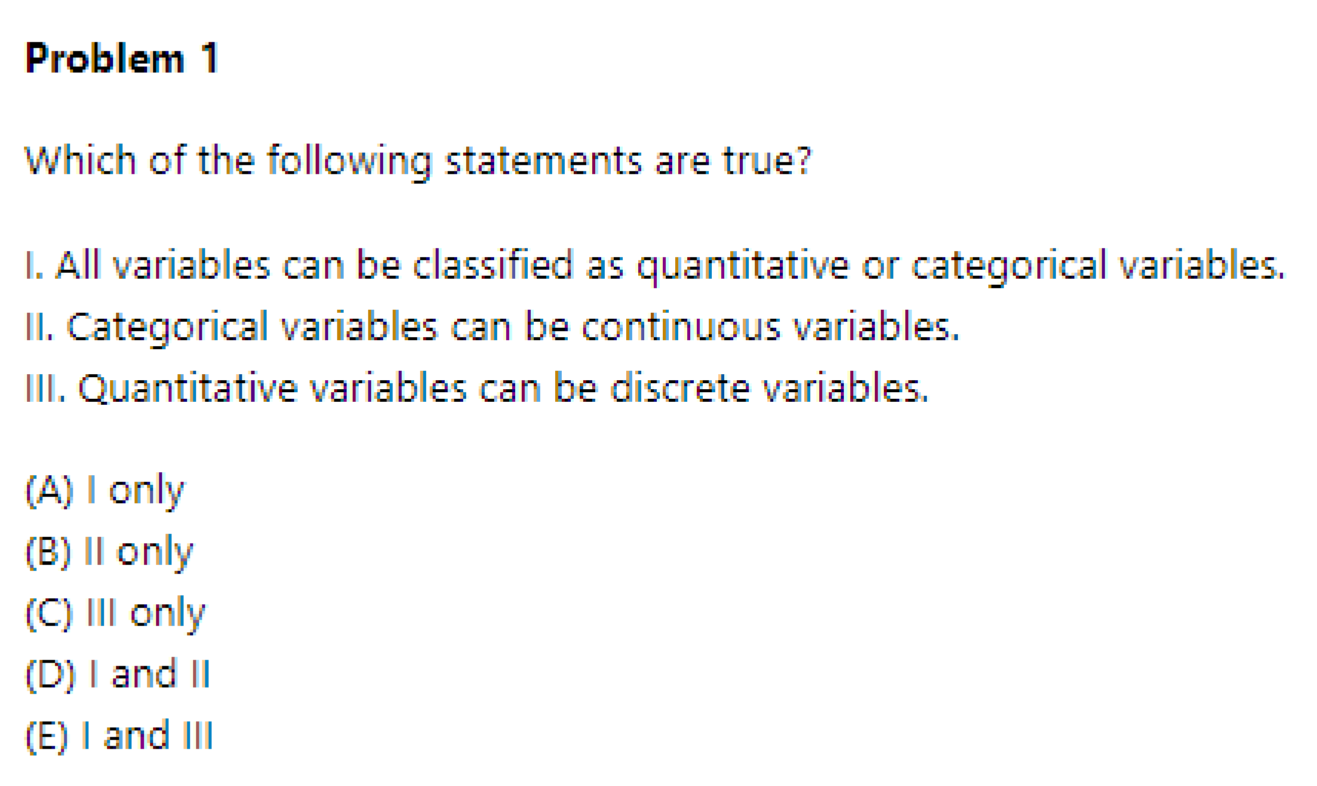
- 1. Cars described as compact, midsize, and full-size.
- 2. Colors of M&M candies.
- 3. Weights of M&M candies
- 5. types of markers (washable, permanent, etc.)
- 6. time it takes to sing the National Anthem
- 7. total annual income for statistics students
- 8. body temperatures of bears in the north pole
- 9. teachers being rated as superior, above average, average, below average, or poor

- 1. Cars described as compact, midsize, and full-size. ordinal
- 2. Colors of M&M candies. nominal
- 3. Weights of M&M candies ratio
- 5. types of markers (washable, permanent, etc.) nominal
- 6. time it takes to sing the National Anthem ratio
- 7. total annual income for statistics students ratio
- 8. body temperatures of bears in the north pole interval
- 9. teachers being rated as superior, above average, average, below average, or poor ordinal

QUIZ II. Identify as Qualitative or Quantitative. If it is quantitative, tell if it is discrete or continuous

- 10. hair color of the math teachers at PHS
- 11. the number of people that prefer Pepsi over Coke
- 12. the weight of your sister’s car (in pounds)
- 13. the number of criminal indictments against Michael Vick
- 14. the length of his jail sentence
- 15. your telephone area code
- 16. how fast you were going when you were pulled over for speeding down Main Street (in MPH)
- 17. the way you felt when you were pulled over for speeding down Main Street

- 10. hair color of the math teachers at PHS qualitative
- 11. the number of people that prefer Pepsi over Coke quantitative discrete
- 12. the weight of your sister's car (in pounds) quantitative continuous
- 13. the number of criminal indictments against Michael Vick quantitative discrete
- 14. the length of his jail sentence quantitative continuous
- 15. your telephone area code qualitative
- 16. how fast you were going when you were pulled over for speeding down Main Street (in MPH) quantitative continuous
- 17. the way you felt when you were pulled over for speeding down Main Street qualitative

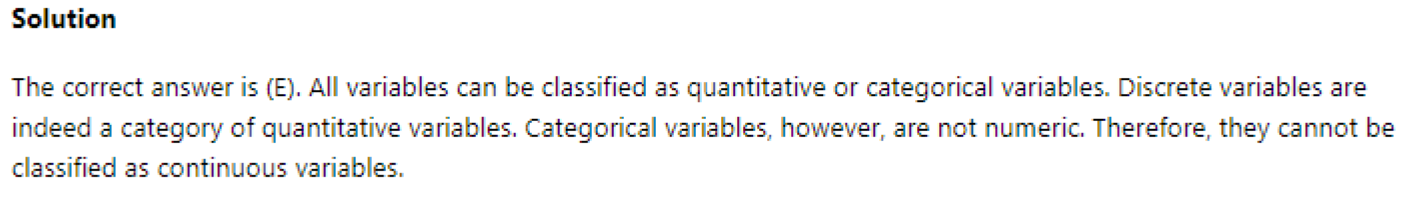


Problem 1

Which of the following statements are true?

I. All variables can be classified as quantitative or categorical variables.
II. Categorical variables can be continuous variables.
III. Quantitative variables can be discrete variables.

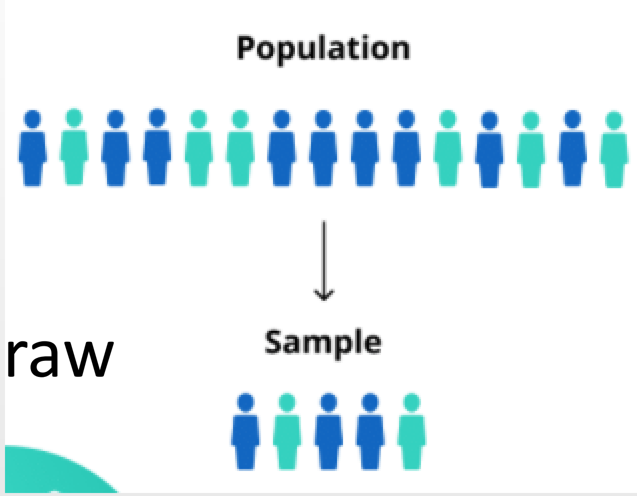
(A) I only
(B) II only
(C) III only
(D) I and II
(E) I and III



Solution

The correct answer is (E). All variables can be classified as quantitative or categorical variables. Discrete variables are indeed a category of quantitative variables. Categorical variables, however, are not numeric. Therefore, they cannot be classified as continuous variables.

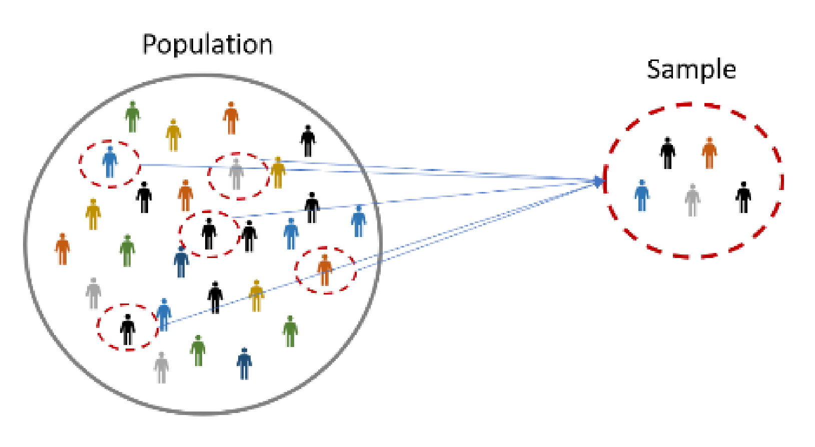
Population



- A **population** is the entire group that you want to draw conclusions about.
- A **sample** is the specific group that you will collect data from.
- The size of the sample is always less than the total size of the population.
- In research, a population doesn’t always refer to people.
- It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc.

Use "population" when:

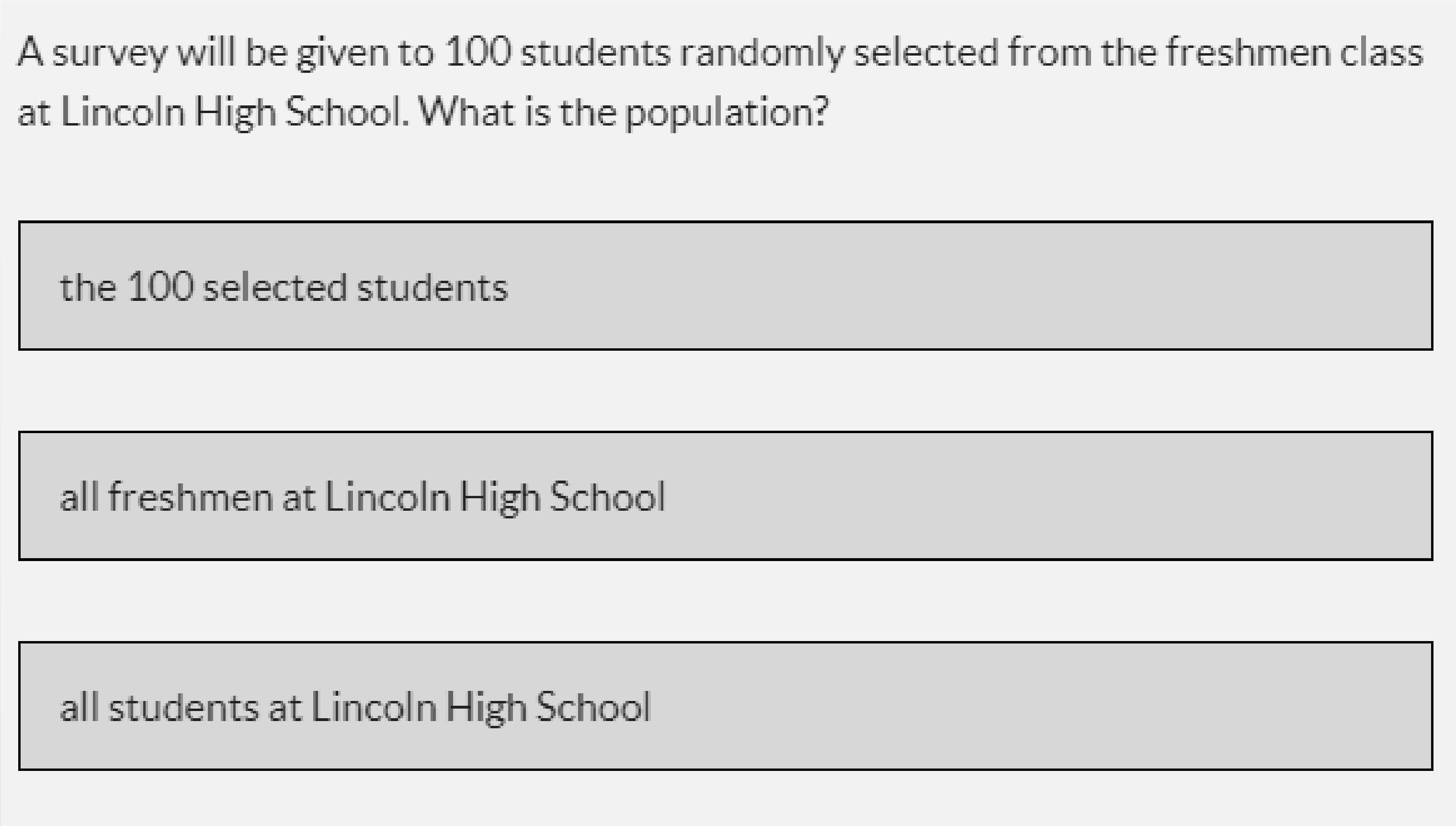
1. you know you have the entire population.
2. you have a sample of a larger population, but you are only interested in this sample (and you will not be generalizing your findings to the entire larger population).



Use "sample" when:

1. you have a sample of a larger population, **and** you wish to generalize your findings from this sample to the entire larger population from which this sample was taken. The sample will be used as an estimate of the population.

Some questions will clearly state whether you are working with a population or a sample. If no statement is present, ask yourself if the statistical findings will be used to describe a larger group. If the answer is yes, you are working with a sample.
Real world statisticians primarily work with sample situations, since real-world data can be overwhelmingly large.



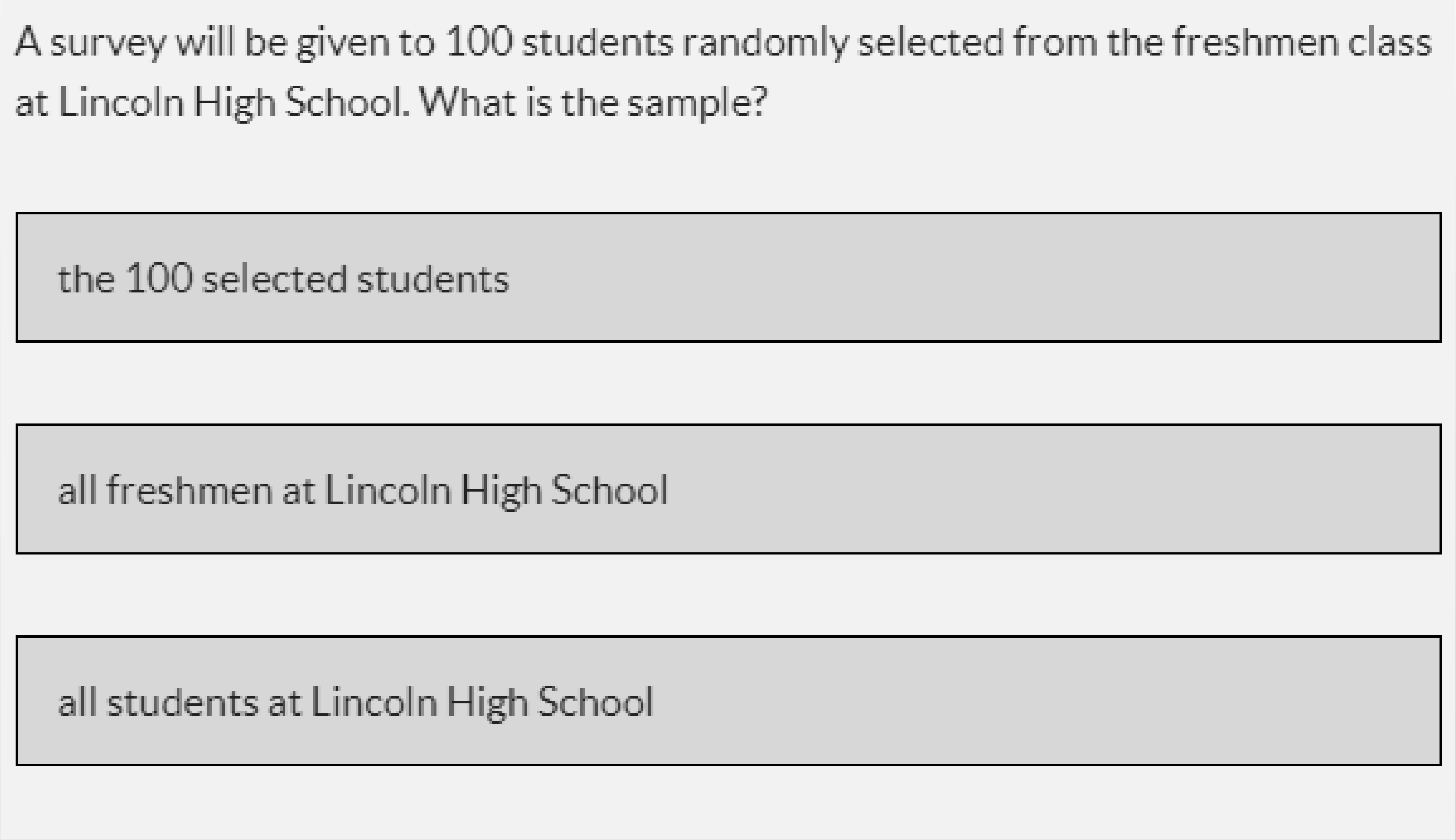
A survey will be given to 100 students randomly selected from the freshmen class at Lincoln High School. What is the population?

the 100 selected students

all freshmen at Lincoln High School

all students at Lincoln High School

- all freshmen at Lincoln High School



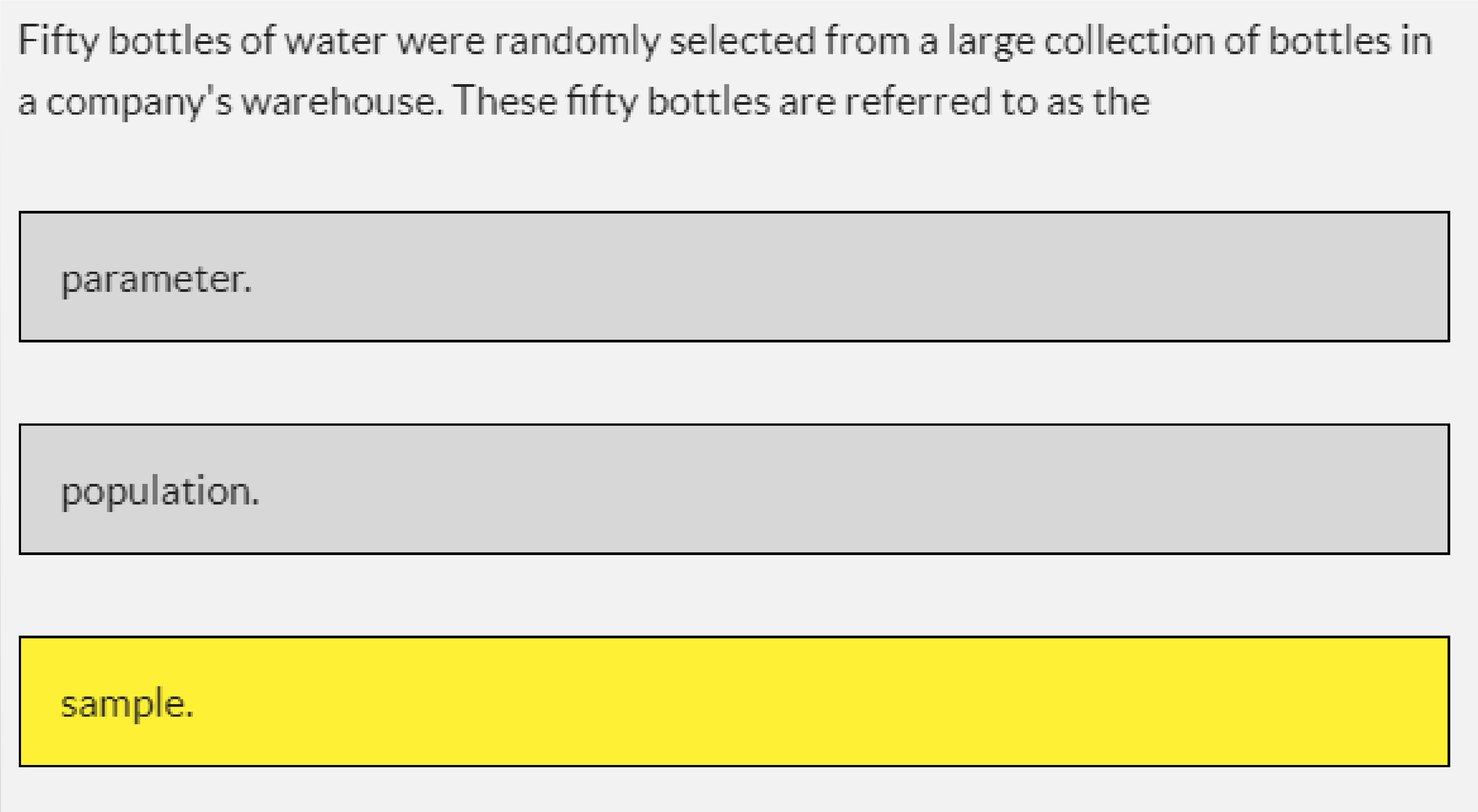
A survey will be given to 100 students randomly selected from the freshmen class at Lincoln High School. What is the sample?

the 100 selected students

all freshmen at Lincoln High School

all students at Lincoln High School

- the 100 selected students



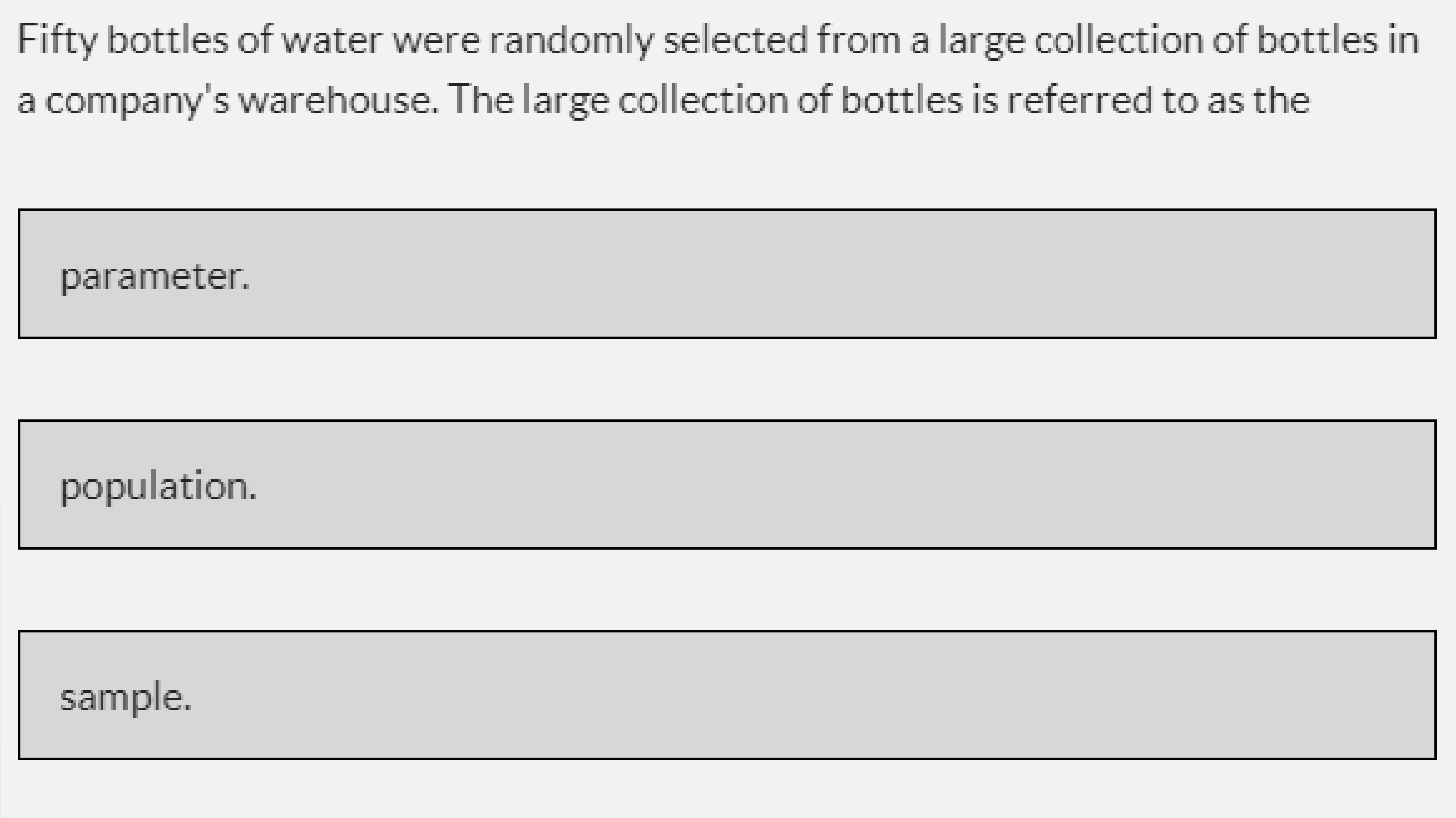
Fifty bottles of water were randomly selected from a large collection of bottles in a company's warehouse. These fifty bottles are referred to as the

parameter.

population.

[✓ sample.]

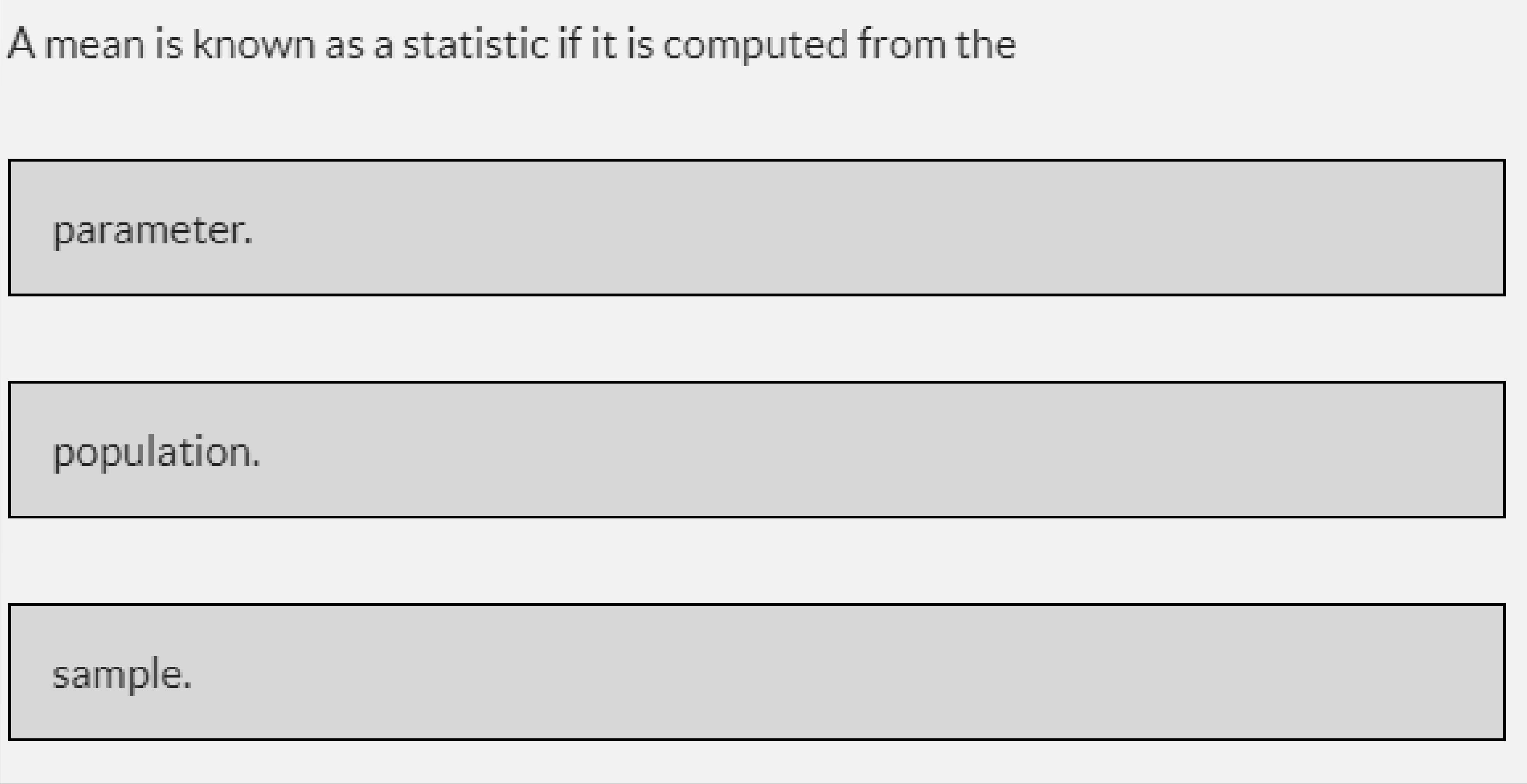
- sample



Fifty bottles of water were randomly selected from a large collection of bottles in a company's warehouse. The large collection of bottles is referred to as the

- parameter.
- population.
- sample.

- population.



A mean is known as a statistic if it is computed from the

parameter.

population.

sample.

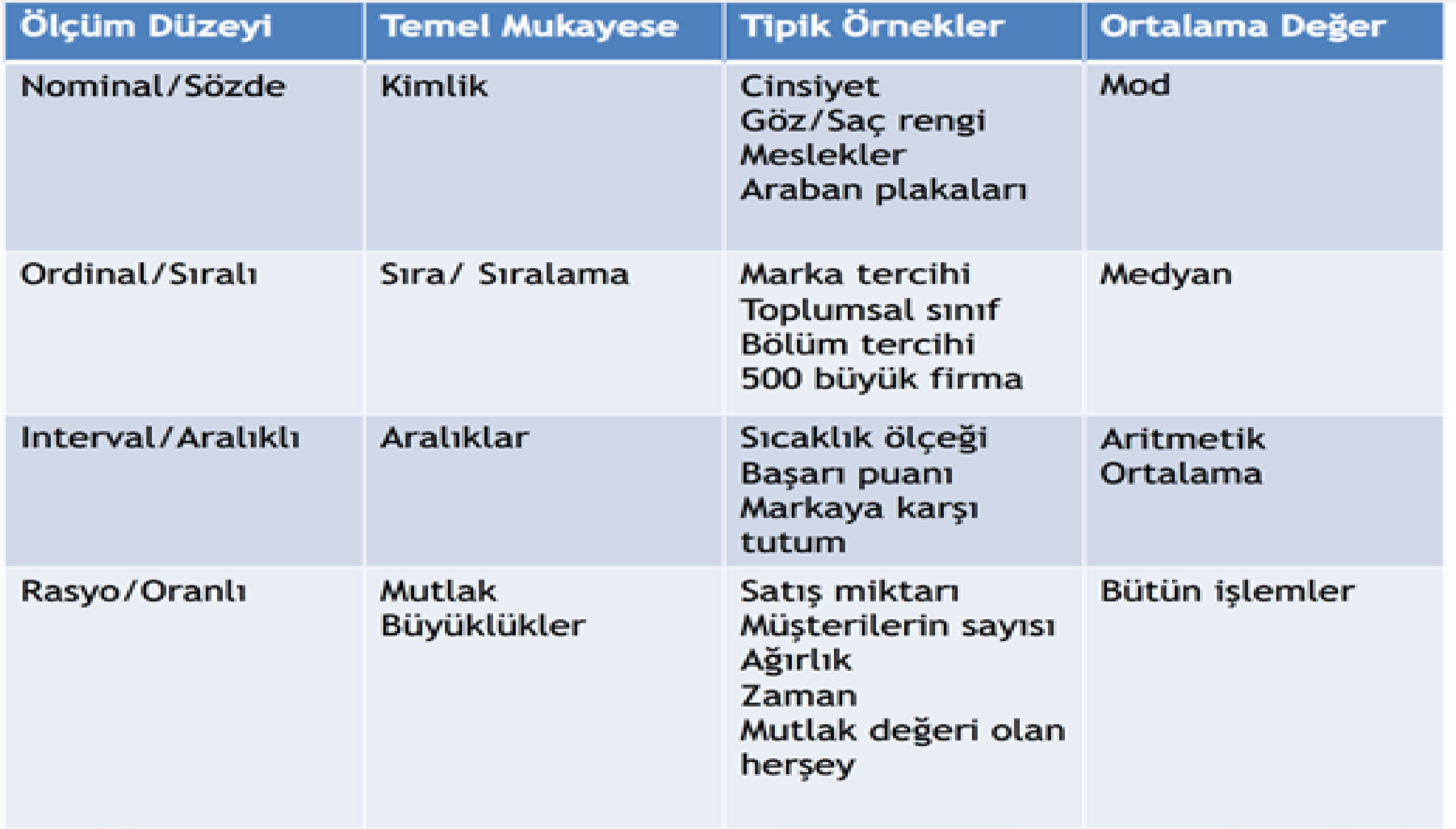
- sample.

Ölçme ve Ölçek (ölçüm) düzeyleri

- Nominal (sözde)
- Ordinal (Sıralı)
- İnterval (Aralıklı)
- Ratio (Oransal)

Sözde ölçekten oranlı ölçeğe gidildikçe ölçeğin ölçme gücü (verilerin kalite düzeyi) artar. Düşük ölçüm düzeyinde (nominal) elde edilen veriler, araştırmacıya pek fazla analiz imkanı sağlamaz.

Ölçme ve Ölçek (ölçüm) düzeyleri



Ölçüm Düzeyi	Temel Mukayese	Tipik Örnekler	Ortalama Değer
Nominal/Sözde	Kimlik	Cinsiyet Göz/Saç rengi Meslekler Araban plakaları	Mod
Ordinal/Sıralı	Sıra/ Sıralama	Marka tercihi Toplumsal sınıf Bölüm tercihi 500 büyük firma	Medyan
Interval/Aralıklı	Aralıklar	Sıcaklık ölçeği Başarı puanı Markaya karşı tutum	Aritmetik Ortalama
Rasyo/Oranlı	Mutlak Büyüklükler	Satış miktarı Müşterilerin sayısı Ağırlık Zaman Mutlak değeri olan herşey	Bütün işlemler

Ölçek düzeyleri

Ölçekler	
Nominal (sınıflayıcı - kategorik)	Sayılar-nesneler arasında bir ilişki öngörülür, nesnelerin sadece gruplandırıldığı. Sayılar ve harfler sadece bir kimliktir. Bir üstünlük ifade etmez. Sadece frekans dağılımı yapılabilir. TC kimlik No herkeste vardır ve nicel değil nitel büyüklüklerdir.
Ordinal	Bir nesnenin belirli bir özelliği az mı yoksa çok mu taşıdığı söylenebilir. Nesneler arası farkın mutlak boyutu söylenemez. Ölçülmek istenen şeyler arasındaki sırayı belirler. Eğitim durumu (ilköğretim, lise, üniversite, lisansüstü şeklindedir.) Tercih edilen ilk üç spor ayakkabı markası…. (sıralama var ama fark bilinememektedir.)
Interval	eşit aralıklı ölçüm düzeyinde nesnelerin sıralanmasında sayısal olarak değerlendirilir. Mutlak sıfır noktasına sahiptir. 1-2 ve 6-7 arası fark eşit sayılmaktadır. (Likert, semantik farklılıklar, ölçeği) Ordinal ölçeğin tüm özelliklerini taşır.
Oranlı	Daha önceki üç ölçeğin özelliklerini taşımakla birlikte, mutlak sıfır noktası vardır. Büyüklükler arası bir oran söz konusudu. Kaç çocuğunuz var? Yaşadığınız şehirde kaç süpermarket var? Yıllık cironuz ne kadardır?