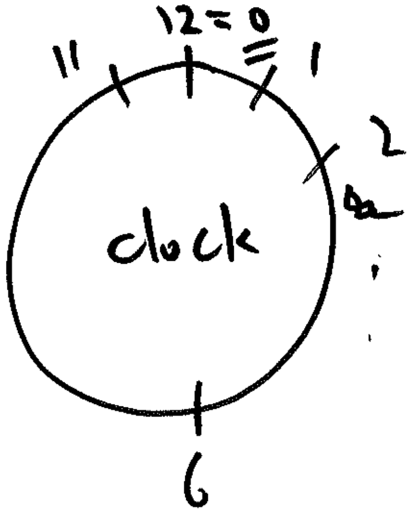
STAT 7
3 Oct 19



what time?

quant

cont.

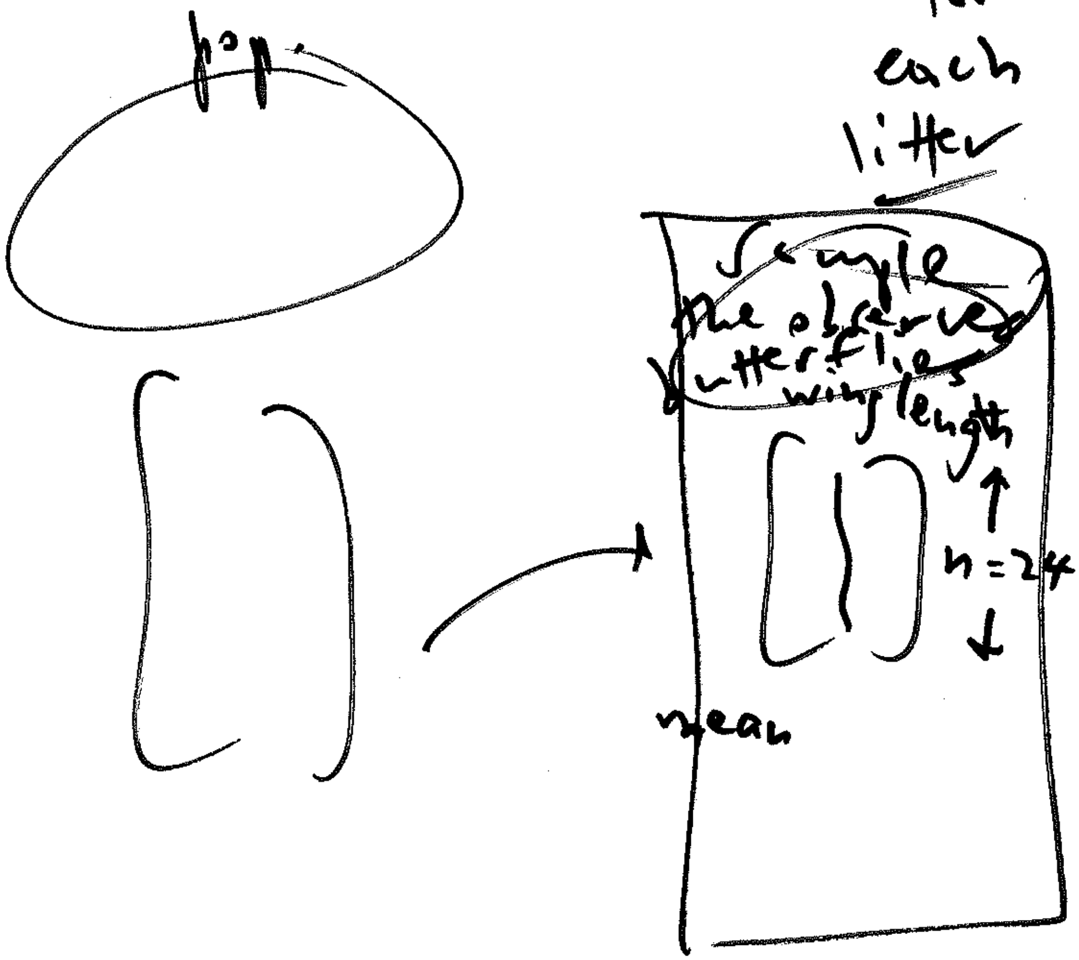
interval

litter size

6
3
~~2.27~~
~~0~~
$\boxed{0}$

quant.
disc.
ratio.

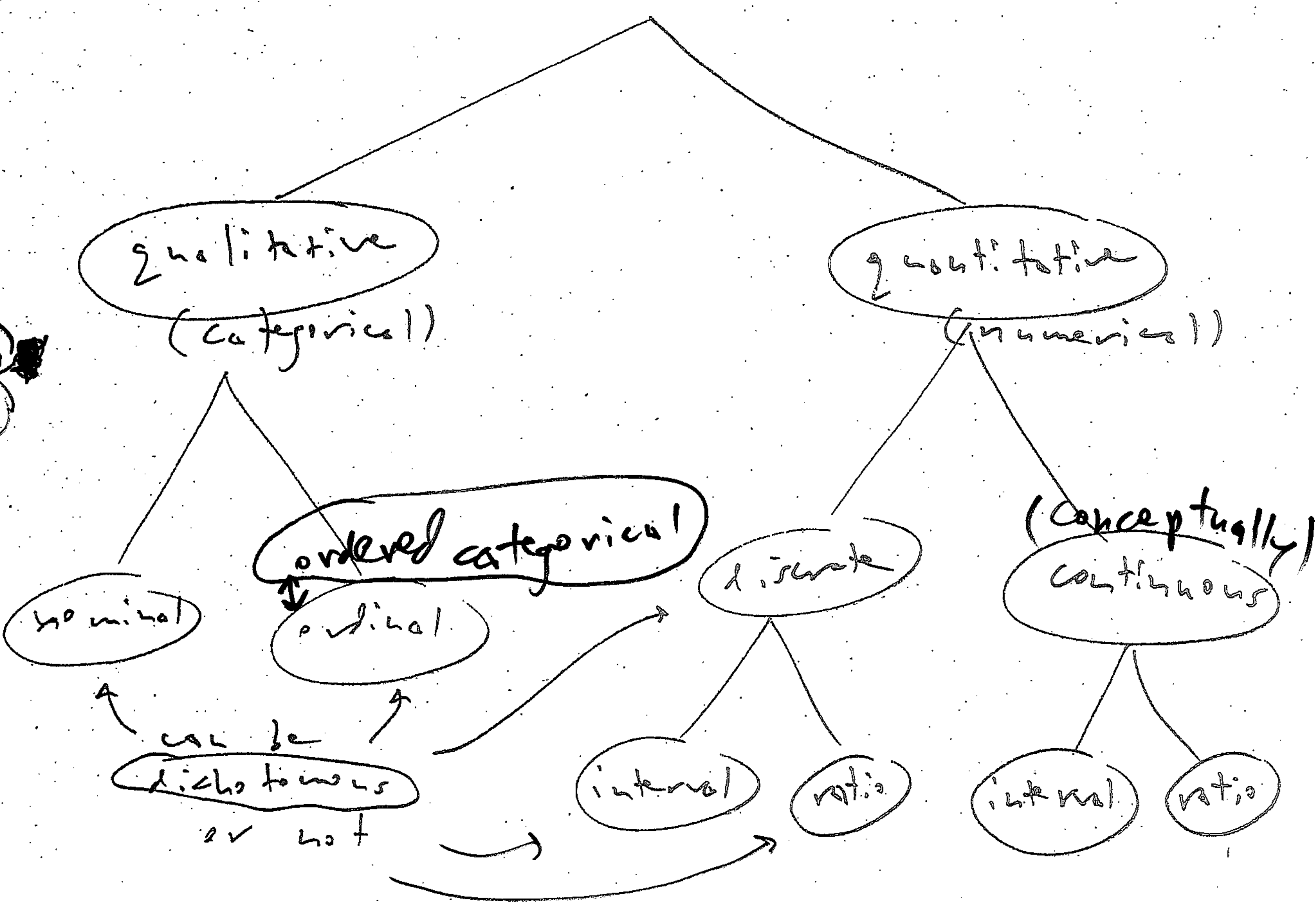
1 row for each litter



graphical / numerical

description of existing data set

Variable types



R-[redacted] 21

1.3 Descriptive Methods

As we'll soon discuss, it's sometimes both **useful** and **meaningful** to **summarize** a variable by taking its **mean** (just add 'em up and divide by how many there are); the computer has done this for us in the table above in the **final column**.

The problem is, of course, that the **mean** is **meaningful** only for the **age** variable (because it's **quantitative** [ratio, discrete]; the other two variables are **qualitative** [nominal]).

The point: The **right way** to **analyze** a variable often depends on the **scale** on which it's measured.

1.3.1 Graphical descriptive methods. **Example: butterfly wing lengths**. Zar (1999) gives data from a sample of $n = 24$ immature monarch butterflies, in which the **variable** of interest (we might call it y; T&T would call it x) is **wing length** (in cm):

4.4 3.6 4.1 3.3 3.5 3.8 4.5 4.3 4.3 4.0 4.1 3.6
4.0 4.0 3.8 3.8 3.9 4.2 4.2 4.1 3.7 3.9 4.0 3.9

(This is just **shorthand** for a **data set** with $n = 24$ **rows** (**subjects** = **butterflies**) and 1 **column** (**variable** = **wing length**), written in this manner to save space.)

How might we **summarize** this variable in a way that would allow us to **see patterns** (**graphical** summaries) and to capture **most of the information it contains** in **fewer than 24 numbers** (**numerical** summaries)?

Raw Frequency Distribution

As long as the **order** in which the data values were listed above is **not relevant**, the first step would be to **sort** the data from **smallest** to **largest**:

3.3 3.5 3.6 3.6 3.7 3.8 3.8 3.8 3.9 3.9 3.9 4.0
4.0 4.0 4.0 4.1 4.1 4.1 4.2 4.2 4.3 4.3 4.4 4.5

Now we can see that there are a number of **duplicate values** (caused by **rounding** the wing length measurement to the nearest cm).

This suggests a **further summary** in which we keep track of the **values** of the variable and the **raw frequencies** (the numbers of times those values are attained):

Value	Frequency
3.3	1
3.4	0
3.5	1
3.6	2
3.7	1
3.8	3
3.9	3
4.0	4
4.1	3
4.2	2
4.3	2
4.4	1
4.5	1
Total	$n = 24$

This is called a **raw frequency distribution** (or **frequency table**) for the variable y (sometimes people just refer to the **distribution** of y, or ask "How is y **distributed**?").

Raw Frequency Histogram

The **table** on the previous page is not as easy to **absorb** as it would be if we could display it **graphically**.

Since it has **two columns** or **dimensions**, it's natural to make a plot in which one dimension (**horizontal**, say) is the **values** the variable takes on and the other (**vertical**, say) is the **raw frequencies** — the resulting graph is a (raw frequency) **histogram** of the variable y:

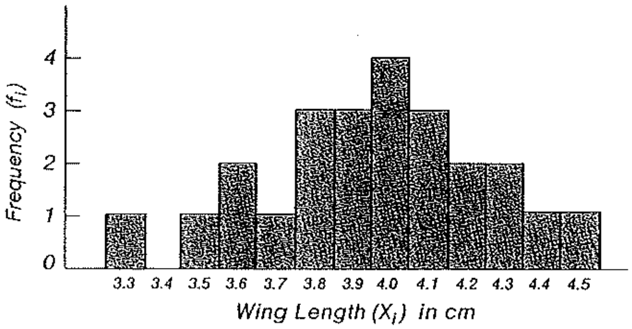


Figure 3.1 A histogram of the data in Example 3.2. The mean (3.96 cm) is the center of gravity of the histogram, and the median (3.975 cm) divides the histogram into two equal areas.

A **histogram** is a special case of a **bar graph**: a plot with **locations** identified along the **horizontal** axis corresponding to **values** a variable takes on and **bars** over those locations with **heights** give by the (raw) **frequencies** of those values.

A **bar graph** can be drawn as a summary of **any qualitative (nominal or ordinal) variable**; there is no unique place called "yes" or "red" on the number line, but you can just **invent arbitrary horizontal locations** and make a useful plot anyway.

Strictly speaking, what makes a **histogram** a histogram is that the variable in question is **quantitative** (so that the values do have unique locations on the number line) — histograms can be made for either **discrete** or **continuous** variables.

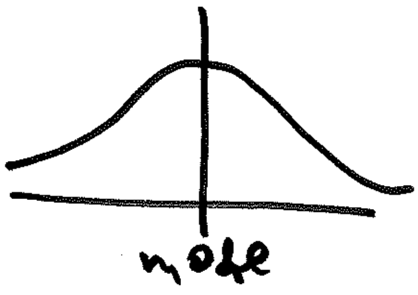
y = wing length (cm) (order irrelevant) judgment

$$\begin{bmatrix} 4.4 = y_1 \\ 3.6 = y_2 \\ 4.1 \quad \cdot \\ 3.3 \quad \vdots \\ \vdots \\ 3.9 = y_{24} = y_n \end{bmatrix} \quad n = 24 \quad \xrightarrow{\text{sort}} \text{from smallest to largest} \quad \begin{bmatrix} y \\ 3.3 \\ 3.5 \\ 3.6 \\ 3.6 \\ 3.7 \\ \vdots \\ 4.5 \end{bmatrix}$$

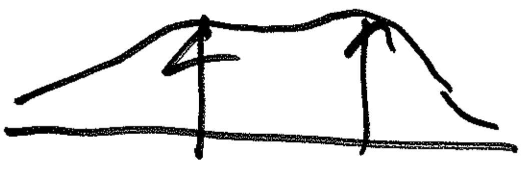
identical information

value	count (raw frequency)
3.3	1
3.4	0
3.5	1
3.6	2
4.0	$\boxed{4}$
4.5	1

highest frequency = mode



bimodal

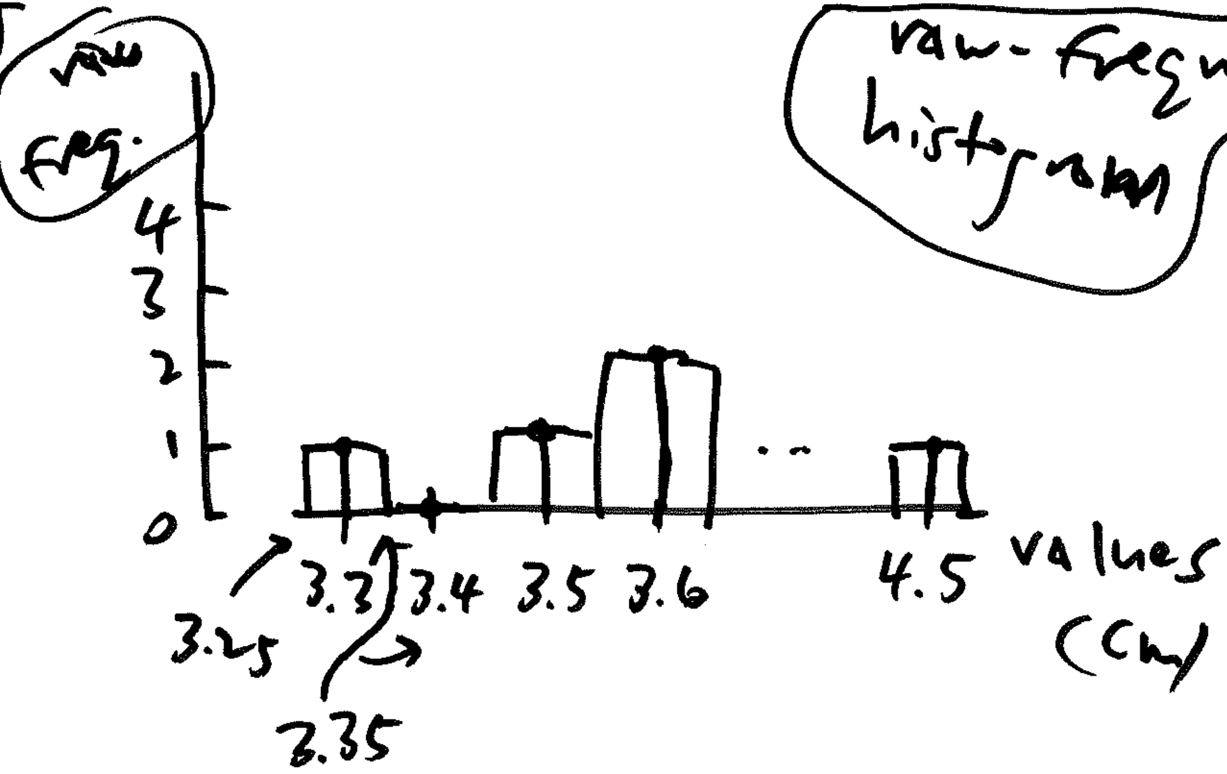


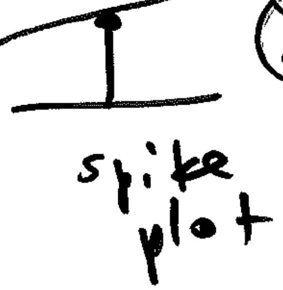
height of adults

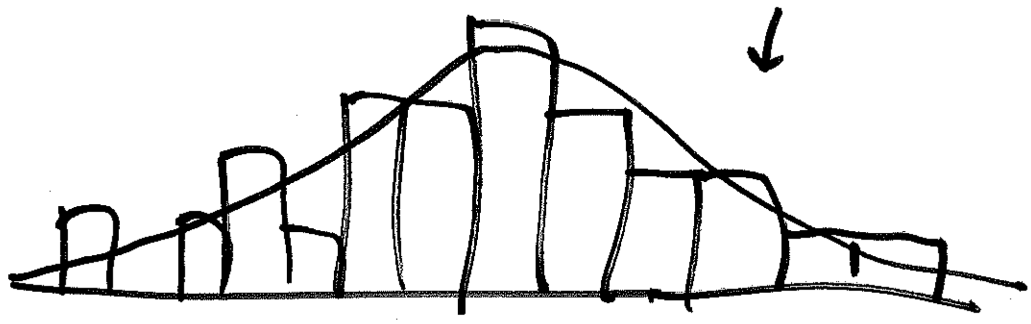
(>1) → (multi)modal

raw-frequency histogram

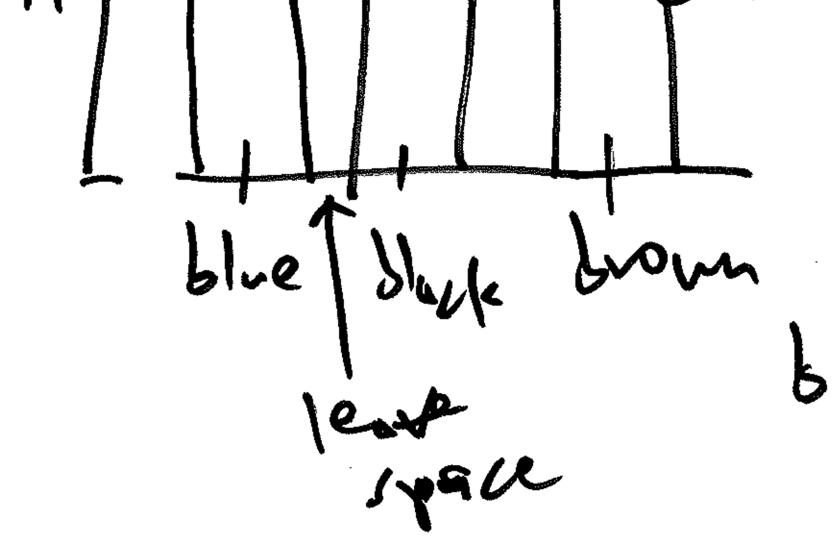
3







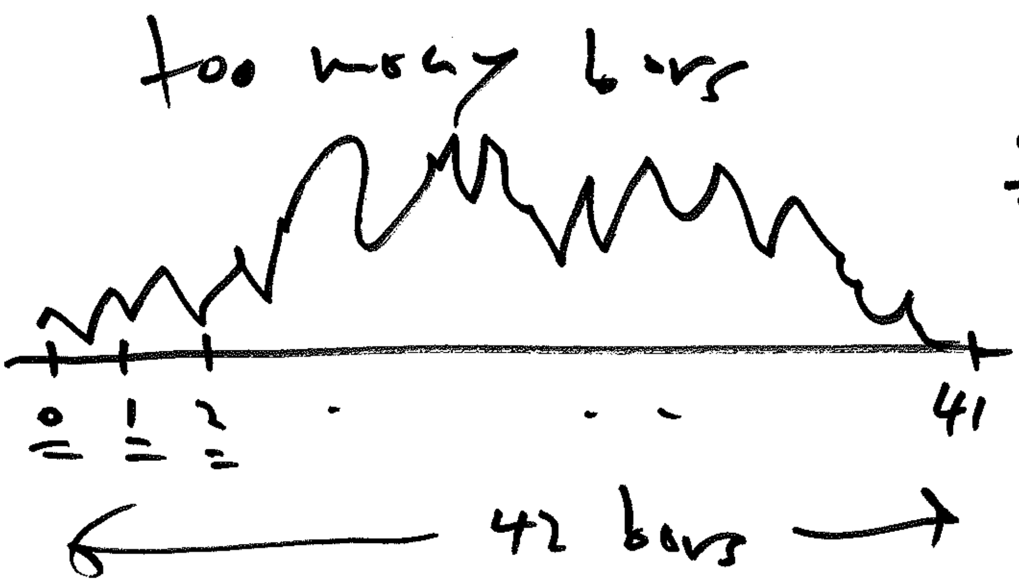
2nd



blue
black
brown

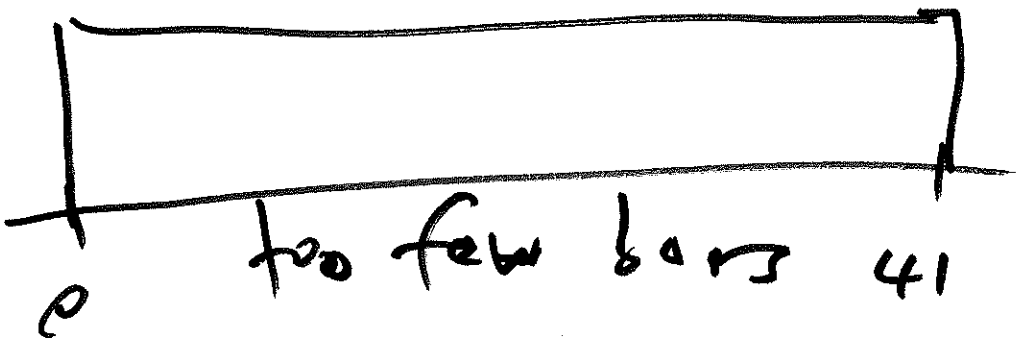
blue	14
black	19
brown	28
	61

bar graph? yes



bad = a hist. ^ lost basic shape in noise

n = 424



bad. = a hist. ^ lost all sense of <u>shape</u> of dist.

More Graphical Examples

vines
ltb
vine
⋮

1 row for each nest

qual
nominal
hist? no
bar graph? yes
dich? no

EXAMPLE 1.1 The location of sparrow nests. A frequency table of nominal data.

Nest site	*Number of nests observed*
A. Vines	56
B. Building eaves	60
C. Low tree branches	46
D. Tree and building cavities	49

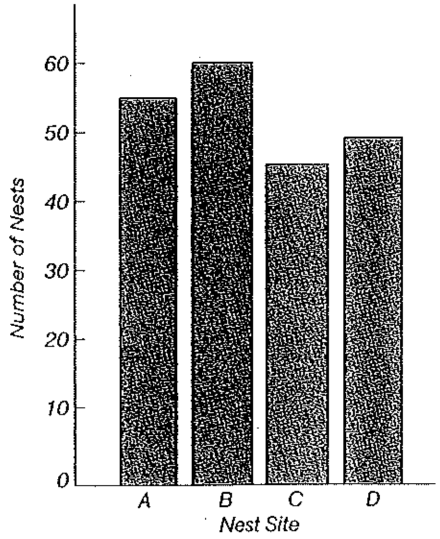


Figure 1.1 A bar graph of the sparrow nest data of Example 1.1. An example of a bar graph for nominal data.

2
4
2
⋮

1 row for each sunfish

qual.
ordinal
dich? no
hist? no
bar graph? yes
dich? no

EXAMPLE 1.2 Numbers of sunfish, tabulated according to amount of black pigmentation. A frequency table of ordinal data.

Pigmentation class	*Amount of pigmentation*	*Number of fish*
0	No black pigmentation	13
1	Faintly speckled	68
2	Moderately speckled	44
3	Heavily speckled	21
4	Solid black pigmentation	8

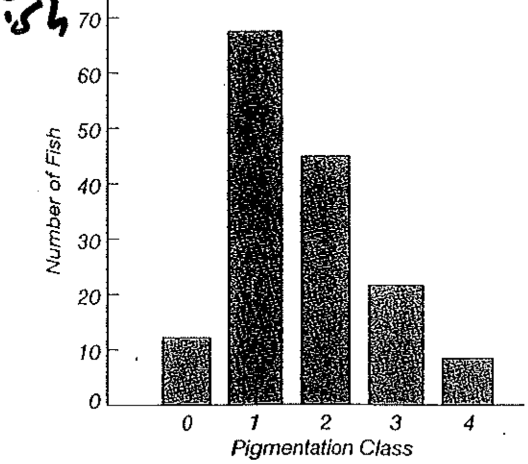


Figure 1.3 A bar graph of the sunfish pigmentation data of Example 1.2. An example of a bar graph for ordinal data.

L- 18

Graphical Examples (continued)

EXAMPLE 1.3 **Frequency of occurrence of various litter sizes in foxes. A frequency table of discrete, ratio-scale data.**

Litter size	*Frequency*
3	10
4	27
5	22
6	4
7	1

$\begin{bmatrix} 5 \\ 3 \\ \vdots \end{bmatrix}$ 1 row for each litter

quant.
disc.
ratio
dich? no
hist? yes

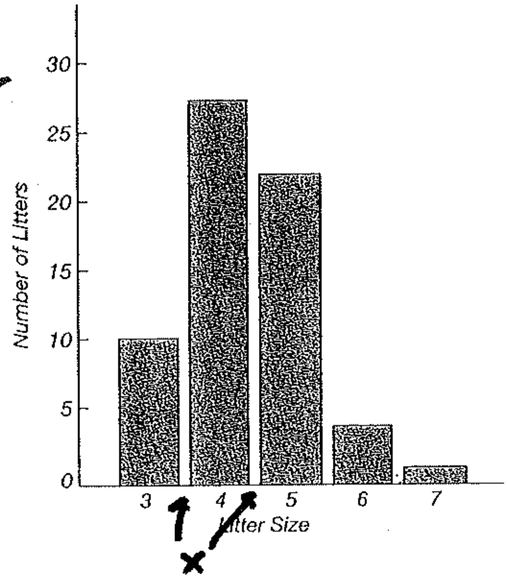


Figure 1.4 A bar graph of the fox litter data of Example 1.3. An example of a bar graph for discrete, ratio-scale data.

EXAMPLE 1.4a **Number of aphids observed per clover plant. A frequency table of discrete, ratio-scale data.**

Number of aphids on a plant	*Number of plants observed*	*Number of aphids on a plant*	*Number of plants observed*
0	3	20	17
1	1	21	18
2	1	22	23 ←
3	1	23	17
4	2	24	19
5	3	25	18
6	5	26	19
7	7	27	21 ←
8	8	28	18
9	11	29	13
10	10	30	10
11	11	31	14
12	13	32	9
13	12	33	10
14	16	34	8
15	13	35	5
16	14	36	4
17	16	37	1
18	15	38	2
19	14	39	1
		40	0
		41	1

Total number of observations = 424

#aphids
$\begin{bmatrix} 27 \\ 16 \\ \vdots \end{bmatrix}$ 1 row for each clover plant

quant.
disc.
ratio
dich?
hist?
yes