STAT_1301; Lecture 2; Jan. 11, '24

Now in Ch. 2.

Objectives: visualize quantitative data using frequency distribution or relative frequency distribution histograms.

Remarks: Skim Ch. 2

i) Know the "idea" behind histograms.

ii) " how to interpret histogram shapes (e.g. for multiple choice q's).

Motivating Data Set:

Value of baseball teams in 2015

(will post data and figures on Nexus).

X = team value ; $n = 30$ observations.

Let's group the data and count the number of teams whose worth (in millions of $)

is within the following intervals:

Team value Interval	Frequency	Relative Frequency
601 - 1050	16	16/30 = 0.533
1051 - 1500	9	9/30 = 0.300
1501 - 1950	1	1/30 = 0.033
1951 - 2400	3	3/30 = 0.100
2401 - 2850	0	0/3 = 0
2851 - 3300	1	1/30 = 0.033
	30	≈ 1

A table with the 1st and 2nd columns gives the frequency distribution of the data.

A table with the 1st and 3rd columns gives the relative frequency distribution of the data.

Rationale for Summarizing data using these

distributions ? Interpretability.

Frequency Distribution Histogram:

y-axis : frequency.

x-axis : intervals.

Relative Frequency Distribution Histogram:

y-axis : relative frequency

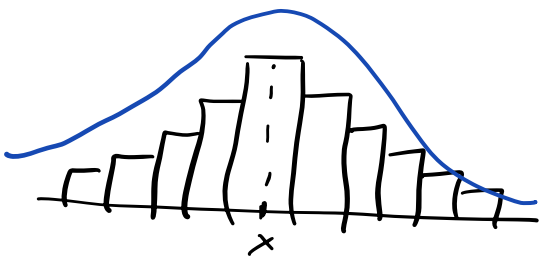
x-axis : intervals

See Nexus "Lecture 2" for figures.

§ 2.2.7 Shapes of Histograms

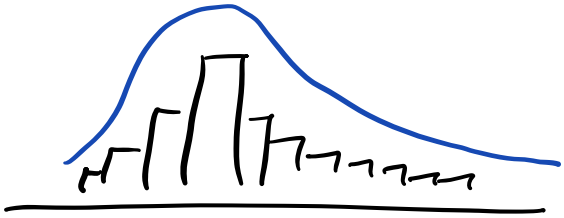
(Know MCQ's)

[Fig. 2.10 in text]



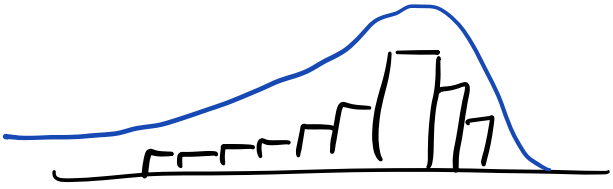
Symmetric

e.g. test scores, IQ
height, weight



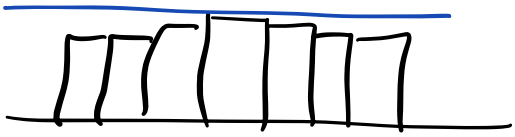
Skewed to the direction (skewness towards "unusual" obs'ns.)

e.g. income

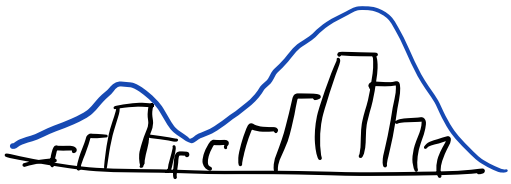


Skewed to the left

e.g. age at retirement



Uniform



bimodal
~ 2 peaks.

Now in Ch. 3
§. 3.1 Measures of Centre for Ungrouped

Motivating Example:

"Median after_tax income for most Canadian families in 2004 rose ... Canadian families with two or more people had an estimate median income after taxes of $54,000 ...." (Source: Stats. Can).

Some measures of central tendency for a population are

i) mean

ii) median

iii) mode

Notation: We use μ ("mu") to denote a population mean.

Def'ns:

i) Mean := Centre of mass/gravity

Population: $X_1, \ldots, X_N$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \leftarrow \text{ Pop. mean}$$

Sample: $x_1, \ldots, x_n$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \leftarrow \text{ Sample mean}$$

ii) Median : = the element/member of the population (Sample) where 50% of the pop. (Sample) is less than it (and 50% of the Pop. (Sample) is greater than it).

iii) Mode : the most frequently occurring member in the pop. (Sample).

Ex. Here are birth weights of a Sample of babies born on a certain day at a facility (in lbs.):

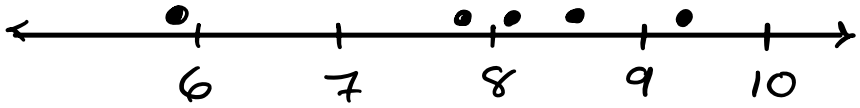
5.9 8.5 9.2 8.1 7.8

The sample mean birthweight is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\sum_{i=1}^{5} x_i = 7.9$$

(Check!)

Dotplot:



Q'n: What is the effect of an unusual observation on $\bar{x}$?

Remove 5.9 from the data. Now the mean birthweight is $\bar{y} = 8.4$.

Message: $\bar{x}$ is sensitive to "unusual obs'ns". Specifically, $\bar{x}$ is pulled in the direction of the extreme observations.

Median: It is the middle value in a data set that has been sorted in increasing order.

Ex. Data: 10 16 −9 −7 0 2 5 7

Find the Sample median.

i) Sort data in increasing order.

−9 −7 0 [2] [5] 7 10 16

median = $\frac{2+5}{2} = 3.5$

ii) Find the location of the location of the middle value.

$\frac{n+1}{2} = \frac{8+1}{2} = 4.5$ ⇝ the location of the median

The Sample median is the average of obs'ns 4 and 5 in the sorted dataset.

Ex. Data: 10 5 19 8 3

Sample median ?

1. Sorted data: Sample median

3 5 **8** 10 19

2. $n = 5$; $\frac{n+1}{2} = \frac{5+1}{2} = 3$ is the location of the median in the sorted data.

Q'n: What is the sample median of the data set 3 5 **8** 10 190 ?

Sample median

MCQ Remark: The median statistic is robust/resistent to "unusual obs'ns".

Mode: = the obs'n(s) that occur(s) with highest frequency.

Ex. Data

21 19 27 22 29 19 25 21

22 ✓ 30

Modes: 19, 21 and 22.

Ex. Data: 21 19 27 22 29

Mode(s) ?

No modes because every obs'n has the same frequency of one.