

Introduction to Statistics.

(1)

Lecture - I.

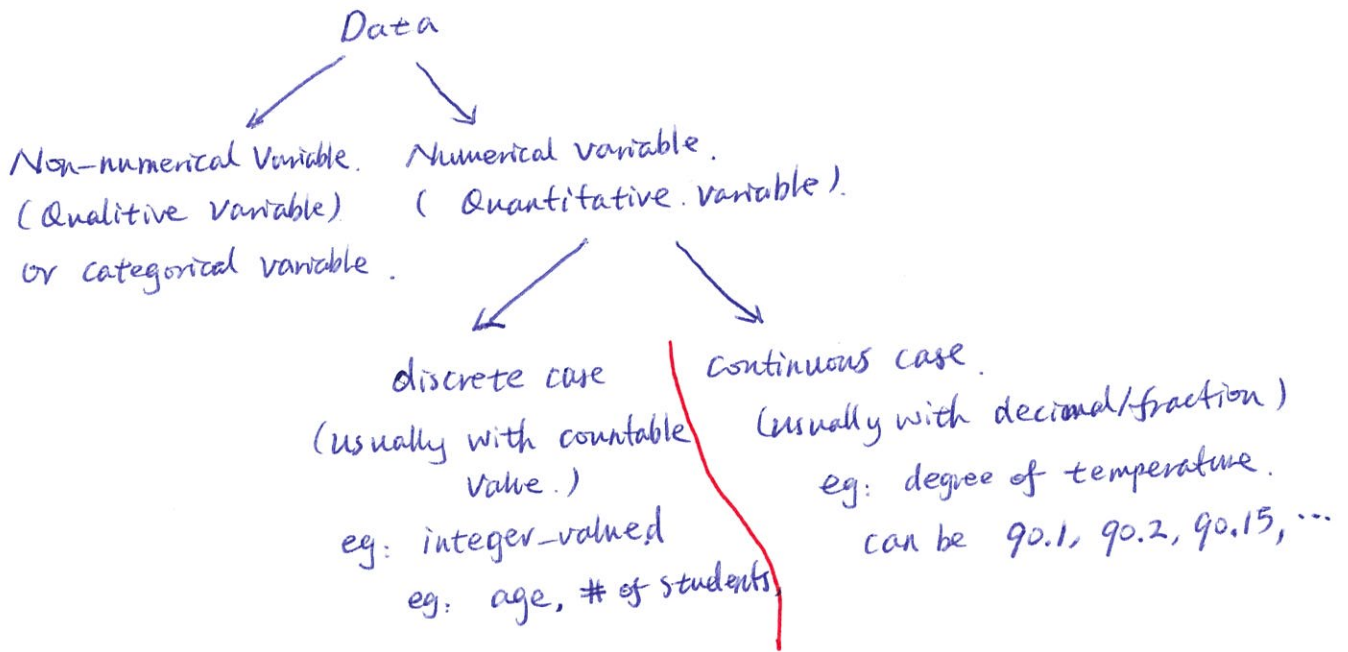
(I): What is Statistics?

⇒ Statistics is the science of conducting studies to collect,
organize, summarize, analyze and draw conclusions
from data.
Middle level stats. *Advanced level stats.* *entry-level stats*

*. What is data? What's inside of a data?

⇒ Data contains variables (column);

Variable is an attributes that can assume different values.



eg:

Studentid	Name	Gender	score
96000123	ABC	M	95
96000456	DEF	F	96
⋮	⋮	⋮	⋮

⇒ This dataset has 4 variables: Name, Gender, score
Categorical variable *Numerical variable*

Note: Studentid can be considered as numerical or Non-numerical, depending on case.

(II) 2 areas of statistics:

2)

a) Descriptive statistics: used to describe a situation.

b) Inferential statistics: make reference from sample to population.

* What is population? consists of all subjects of being studied.

What is sample? \Rightarrow a group of subjects of being studied.

eg: say, we want to study the happiness month of saddleback students.

\Rightarrow : population: all the saddleback college students.

sample: a group of saddleback college students, eg. Math 10 students.

We did a survey for all Math 10 students,

Is December your most happy month of the year?

\Rightarrow : 14 students said yes! $\Rightarrow 14/(14+36) = 28\%$

36 students said No! $\Rightarrow 36/50 = 72\%$

\Rightarrow We can say 28% of students in Math 10 class feel December is the most happy month. } descriptive stats.

If we want to make conclusion: 28% of all saddleback college students feel December is the most happy month. } Inferential statistics.

\Rightarrow Then it would be a referential statistics.

(but it's NOT a good referential statistics)

Since Math 10 student can't represent all college students.

\Rightarrow : To make it good referential statistics,

We need to choose more representative sample!

* Stratified sample: select (randomly) a few students from each class to better represent saddleback college students.

III) Probability:

⇒ probability deal with events that occur by chance.

A lot of applications in gambling/casino business.

insurance business, etc.

⇒ probability is usually used as ^a basis for inferential statistics.

IV) Organizing Data,

eg:

SID	score
id1	100
id2	96
id3	81
id4	92
id5	72
id6	61
id7	50
id8	85
id9	76
id10	49

Extreme values.

⇒ { Highest score : 100
Lowest score : 49

Range: Highest - Lowest
= $100 - 49 = 51$.

Average: $\frac{100 + 96 + \dots + 49}{10} = \dots$

say, we were asked to generate

frequency table based on 5 classes.

⇒ How do we determine the lower level and upper level for each class?

1): First use 'range' / # of classes, to get

$$\text{Class width} = \frac{51}{5} = 10.2$$

(round up to 11)

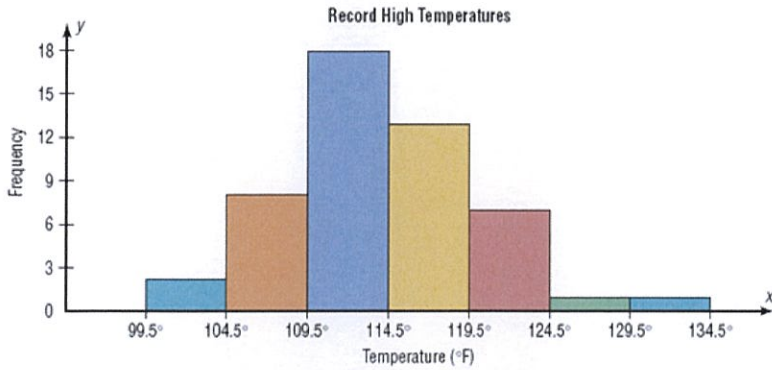
2): The 1st level start with lowest:

	Lower	upper	frequency	%
Level 1	49	49+11=60	2	20%
Level 2	61	71	1	10%
level 3	72	82	3	30%
level 4	83	93	2	20%
level 5	94	100	2	20%
Total:			10	100%

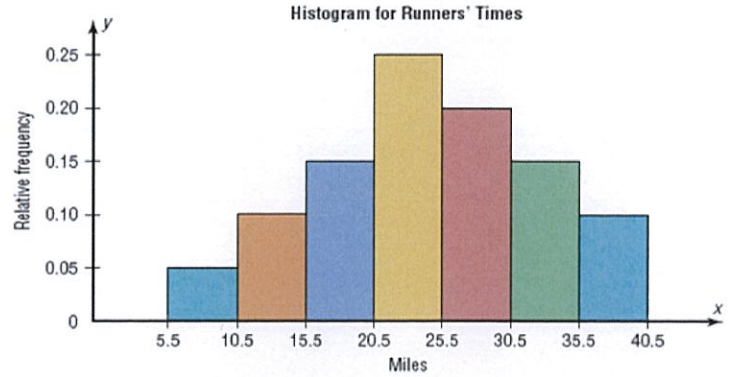
*. General histogram graph, bar chart, pie chart to present data.

Histograms

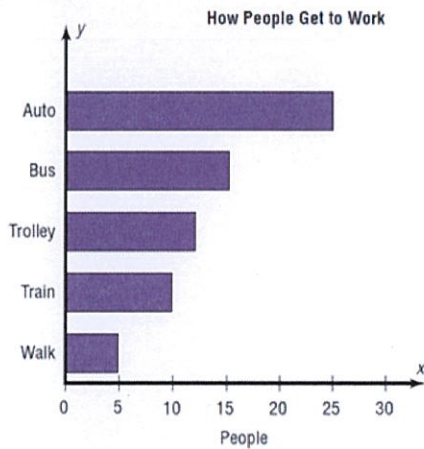
Histograms use class boundaries and frequencies of the classes.



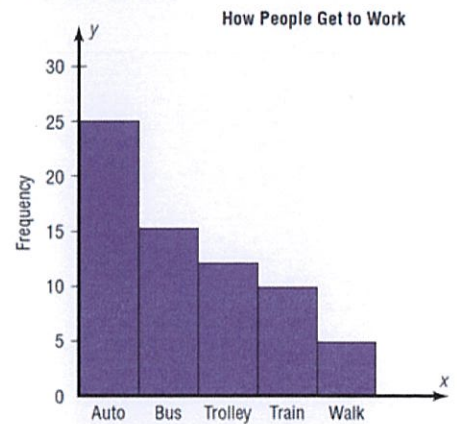
Use the class boundaries and the relative frequencies of the classes.



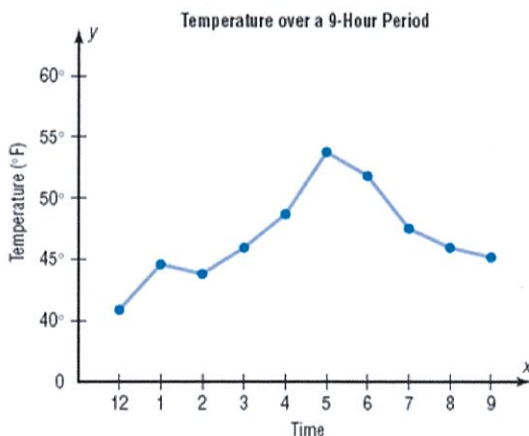
Bar Graphs



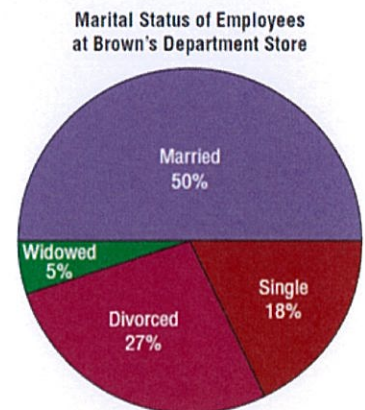
Pareto Charts



Time Series Graphs



Pie Graphs



More Tutorial at www.LittleDumbDoctor.com