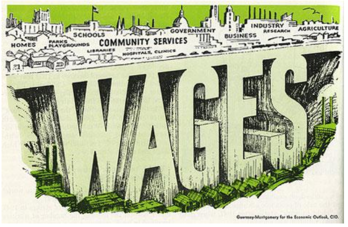
Topic 1: Descriptive Statistics

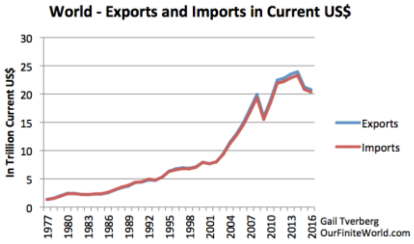
Reference: AWS: Chapters 1 and 2.
Objectives: Basic Statistical Definitions
Methods of Displaying Data

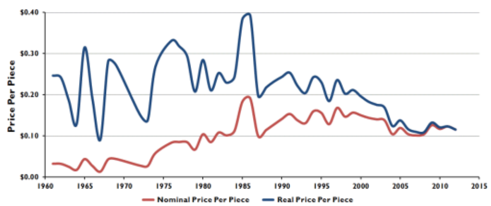
Definitions:

S__________: a numerical piece of information

Example: We are interested in economic data
⇨prices ⇨exports ⇨interest rates
⇨inflation rates ⇨wages







Descriptive Statistics: ways of summarizing or p__________ statistical information efficiently and effectively.

Inferential Statistics – used to assist with decision making when faced with ***un__________***.

■In order to understand the distinction between these definitions, we need to understand the distinction between a "**population**" and a **"sample":**

Population: **All** the ____ items that may be of interest.

Sample: A selected **subset** of the population items
(How should you select this subset?)

The distinction between "population" and sample may depend on context.

<u>Example:</u>

<u>Population</u>: (ALL) New cars sold in Vancouver.

<u>Sample</u>: Ten new cars sold in Vancouver auto____.

▲The statistical sample attempts to provide information that helps us understand some characteristic (__________) of the population.

▲We are trying to infer something about the (general) population from the (_________) sample results.

(Trying to make a generalization about a population, from the results attained from a sample of population.)

Such a process involves **<u>uncertainty. (_____)</u>**

We need to be able to **measure / __________** this, so we can judge the ________ of our inferences.

(Provide a margin of error; numerical measure of ____________;
Population – no uncertainty – have all the facts
Sample – uncertainty – not have the whole picture.)

Our Motivation: Decision making is an essential activity for corporations, government agencies, etc..

Decisions often involve quantitative information.
Such information often involves **uncertainty**.

- *Policy making*
- *Forecasting*

Statistical Inference Involves 3 Basic Procedures:

(1) **E__________** – of population parameter(s) using a sample(s).

Example: minimum price of new car sold
Example: average percent of cups of coffee sold that are decaffeinated.

(2) **H__________ Testing** – testing the validity of some statement about a population.

Example: 10% of all new cars sold are less that $12,000 (Cdn$).
Example: 25% of all coffee sold is decaffeinated.

(3) **F__________g** – Predicting outside the sample.

Example: Minimum price of new car in 2015.
Example: Average amount of coffee sold that is decaffeinated next month.

►*Look at the trends*

Must Learn about:

Dispersion: variance, skew
Central tendency: mean, median and mode

i) **Data characteristics**
ii) **Data ______________** → *Tables and pictures*
iii) **Measuring uncertainty (probability)**
iv) **Tools of statistical __________**

■*Survey design*
■*Assumptions*
■*CL.T.*

Population and Sample Characteristics

Often a population is very _____, so it is useful to summarize its key features by focussing on a few important **characteristics**.

Examples:
"What is the average or most typical population value?"
⇨The average wage of all working Canadians is $54,250.13 per year.

"What range of values does the data cover?"
⇨The number of trucks sold by every Toyota dealership in Canada range from as low as 5 to a maximum of 498 per year.

Such characteristics are called **population __________**.

Numerical Example:

Suppose there are only 10 retail stores in _________ that sell a particular ink cartridge for an old piece of office equipment.

The prices of these cartridges are:

{23.45, 23.23, 20.98, 24.56, 24.05, 23.24, 23.99, 22.99, 25.50, 23.99}

Sum =235.98

The population mean (average) is _______:

$$\mu = mean$$

$$\mu = \frac{1}{N}\left(X_1 + X_2 + \ldots + X_N\right)$$

$$= \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)$$

where:

N = Population size

X_i = i^{th} value

$$\mu = \frac{1}{10}\left(23.45 + 23.23 + \ldots + 23.99\right) = \frac{1}{10}\left(235.98\right) = 23.598$$

The <u>Proportion</u> of values in the population below $__:

$$\Pi = \frac{y}{N} = \frac{2}{10} = \frac{1}{5} = 20\%$$

where:

y = number of values below $23.

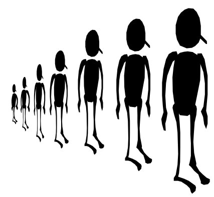
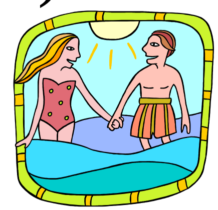
N = Population size.

The most **<u>frequently __________</u>** value in the population is:

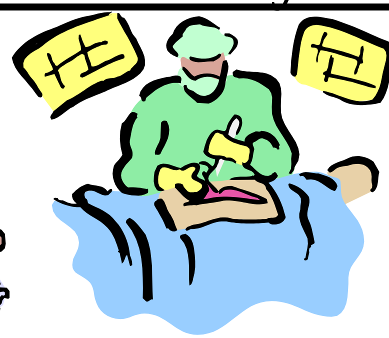
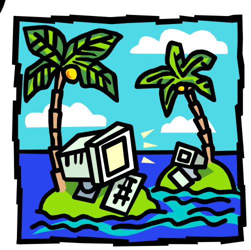
M=$23.99 — occurring twice.

Often we need to work with a ______ of data, instead of the entire population because:

(i) **<u>Population is very _____</u>** - expensive (i.e. labour cost, time.)

(ii) **<u>Part of the population may be in__________</u>**

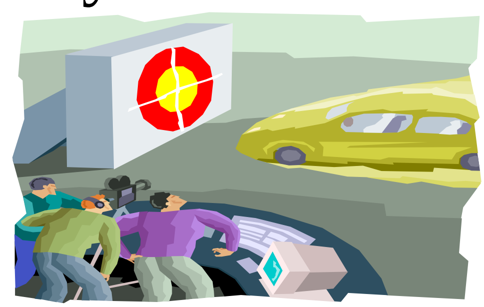




- ☺ Holiday
- ■ Military
- ⌘ Hospital

(iii) **Measurement may be ___________**

Eg. Testing the reliability of an electrical component can only be performed if the component is destroyed; stress test.

Eg. Crash testing ________ for certain safety features.



Eg. Water monitoring for quality control.

⇨The individual sample ______ are called **sample statistics**.

⇨Similarly, any **function** of the sample values is also called a statistic.

⇨The sample **s_______** characterize the feature of a sample in the way that **parameters** characterize a population.

Example: using the ink cartridge data, we choose 3 items (n=3) from the population of 10 prices:

{23.23, 23.99, 20.98}

The **sample mean** price is: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{3}(23.23 + 23.99 + 20.98) = \frac{1}{3}(68.2) = 22.73$

The *<u>sample proportion</u>* ***of prices below $23 is:***

$$p = \frac{y}{n} = \frac{1}{3} = 33.33\%$$

(______ than the population)

There is no most frequently occurring value; each occurs once in this sample.

Comparison of Population and Sample Characteristics:

μ =23.598	$\bar{X}$ = 22.73	______ are different
Π =20%	p=33.33%	proportions are ____________
M=23.99	m	different

Using a sample introduces uncertainty.

(Can sampling error be controlled?) (YES!! _________ ***n**.)*

Data Presentation

(I) Tabular Presentation:

When reporting data, you need to report:

(i) ____

(ii) units of m__________

(iii) method of sampling *(telephone; volunteer)*

(iv) reliability (outliers, rounding)

(v) consistency with other data

(vi) relevance for our purposes

(vii) potential to be _______ (maintained)

Self-rated health, by sex, household population aged 12 and over, Canada, provinces, territories, health regions and peer groups, 2005

Geographic code and name	Self-rated health								
	Total	Very good or excellent		Good		Fair or poor		Not stated	
		number	%	number	%	number	%	number	%
Canada	**27,131,964**	**16,295,063**	**60.1**	**7,781,666**	**28.7**	**3,028,494**	**11.2**	**26,742**	**0.1**
Males	**13,371,912**	**8,097,453**	**60.6**	**3,835,129**	**28.7**	**1,425,163**	**10.7**	**14,167**	**0.1**
Females	**13,760,052**	**8,197,610**	**59.6**	**3,946,536**	**28.7**	**1,603,331**	**11.7**	**12,574 E**	**0.1 E**
10 Newfoundland and Labrador	**448,813**	**288,338**	**64.2**	**106,825**	**23.8**	**53,454**	**11.9**	**F**	**F**
Males	**219,553**	**137,739**	**62.7**	**55,234**	**25.2**	**26,406**	**12.0**	**F**	**F**
Females	**229,259**	**150,599**	**65.7**	**51,591**	**22.5**	**27,049**	**11.8**	**F**	**F**
1011-C Eastern RIHA, N.L.	260,578	171,600	65.9	60,600	23.3	28,255	10.8	F	F
Males	126,614	82,022	64.8	30,666	24.2	13,802	10.9	F	F

1. Health regions are defined by the provincial ministries of health. These are legislated administrative areas in all provinces. The health regions presented in this table are based on boundaries and names in effect as of June 2005. For complete Canadian coverage, each of the northern territories also represents a health region.
2. A "peer group" is a grouping of health regions that have similar social and economic characteristics. The nine peer groups are identified by the letters A through I, which are appended to the health region 4-digit code.
3. In Nova Scotia, zones are aggregations of the nine district health authorities.
4. No data available for "Région du Nunavik" and "Région des Terres Cries de la Baie-James".

Table 1
Most prevalent occupations usually requiring a university degree, women, 1996 and 2016

	1996				2016			
	Workers	Proportion of workers aged 55 and over	Median age	Ratio of younger workers to older workers [2]	Workers	Proportion of workers aged 55 and over	Median age	Ratio of younger workers to older workers [2]
	number	percent	years	ratio	number	percent	years	ratio
Occupation								
Registered nurses and registered psychiatric nurses	214,800	9.6	41.6	4.51	262,500	20.3	42.8	1.56
Elementary school and kindergarten teachers	183,100	7.2	43.7	4.18	238,700	13.8	41.2	2.65
Financial auditors and accountants	52,200	6.4	37.2	4.65	108,400	19.1	43.8	1.35
Secondary school teachers	77,300	7.9	42.7	4.12	94,900	16.0	42.2	2.58
Professional occupations in advertising, marketing and public relations	16,200	5.6	36.7	5.45	60,100	10.9	35.7	4.03
Human resources professionals	14,200	3.7	40.2	10.40	53,200	15.1	40.8	1.82
Other financial officers [1]	12,200	6.9	38.8	3.61	51,600	19.5	44.3	1.21
Social workers	28,400	6.1	39.2	3.97	49,200	16.7	41.4	2.22

A Good Data Table will include:

- ⇨____– what, when where
- ⇨ ______ of measurement
- ⇨ Definitions of symbols / terms
- ⇨Source(s)
- ⇨ Data adjustments – rounding
- ⇨Breaks in the data

There Are Many Potential Pitfalls:

- ▽Misinterpretation of figures (units may differ)
- ▽Misleading ____________
- ▽Mixed reliability (*misinterpretation by collectors*)
- ▽Inadequate _____ / Incomplete title

"– we need a method that will summarize or describe large masses of data without loss or distortion of essential characteristics and make the data easier to interpret. One such method is the arrangement of data into what is called a __________ distribution:"

Frequency Distributions:

A convenient way of summarizing a large set of _______ data.

–Divide values into intervals and report the frequency of o_________ of values in each interval. *(Group by frequency of occurrence)*

"To construct a frequency distribution, it is first necessary to divide data into a limited amount of classes and report the number of times (frequency) an observation falls (is distributed) in to ____ class."

Example: Suppose we have a population of 20 prices:

Class (i)	**Range ($) (width =5)**	**Frequency** (f_i)	**Relative Frequency** $\left(f_i/N\right)$
1	**10 $\leq$ X<15**	**8**	**0.40**
2	**15 $\leq$ X<20**	**4**	**0.20**
3	**20 $\leq$ X<25**	**5**	**0.25**
4	**25 $\leq$ X<30**	**2**	**0.10**
5	**30 $\leq$ X<35**	**1**	**0.05**
		N=20=$\sum f_i$	**1.00**

Relative frequency is the frequency in each class ________ to the total number of observations.

The relative frequency is determined by dividing the frequency of each class by the total number of observations and expressing the result as a _______.

Note: ***with this example, data is in interval form instead of individual observations:***

- ☐ Individual data details are "______"
- ☐ Intervals have equal width – 5 units
- ☐ Intervals are non-overlapping
- ☐ Interval widths are sensible for the data
- ☐ Number of intervals are sensible
- ☐ Intervals are 'closed'
- ☐ Could use ___-______ as representative *(for calculations)*

Also useful to construct a** <u>cumulative frequency distribution</u> **or a** <u>cumulative relative frequency distribution</u>**:

Class (i)	**Range ($)**	*fi*	$\sum f_i$	$\left(f_i/N \right)$	$\sum \left(f_i/N \right)$
1	$10 \leq X < 15$	**8**		**0.40**	
2	$15 \leq X < 20$	***4***		**0.20**	
3	$20 \leq X < 25$	***5***		**0.25**	
4	$25 \leq X < 30$	***2***		**0.10**	
5	$30 \leq X < 35$	**1**		**0.05**	

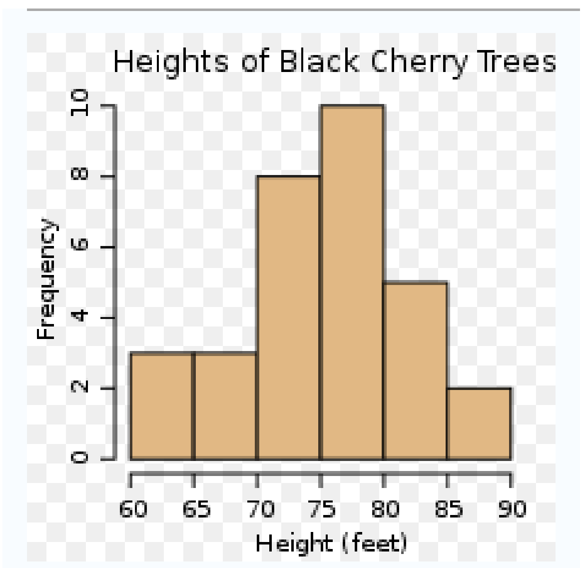
The cumulative frequency is the sum of the absolute frequencies from lowest class to the highest class.

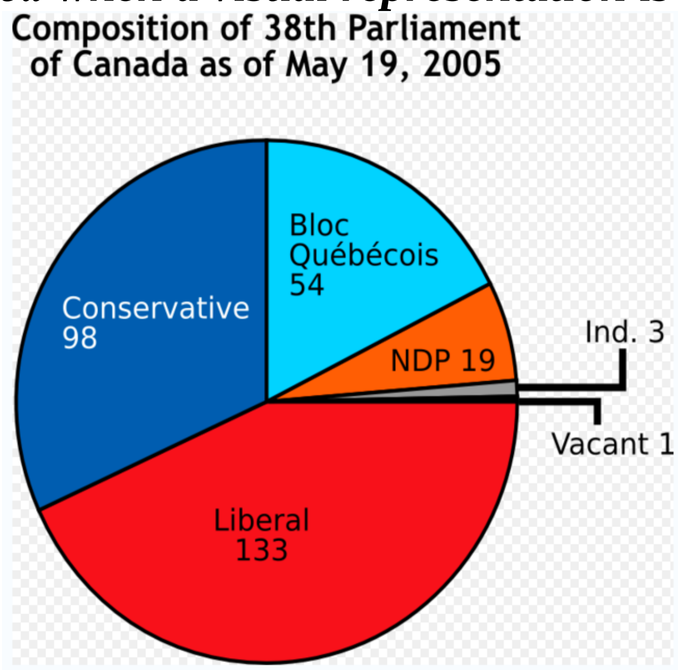
Relative frequency sums to 1.

(B) **Graphical Presentation:**

A graph is another way to summarize data.
More effective if data features are complex.
–i.e. greater impact/ more efficient

"Graphs and charts are usually employed when a visual representation is desired."





However, there is a greater potential for mis-interpretation.

Along with the previous requirements for a good data table, we also need these:

- ▲ All ____ must be labelled
- ▲ _______(s) must be labelled
- ▲ A clear, uncluttered image

Easy to construct graphs corresponding to frequency, relative frequency, cumulative frequency and cumulative relative frequency.

"While it is often useful to arrange the values in a data set into a frequency distribution, many analysts prefer a pictorial presentation."

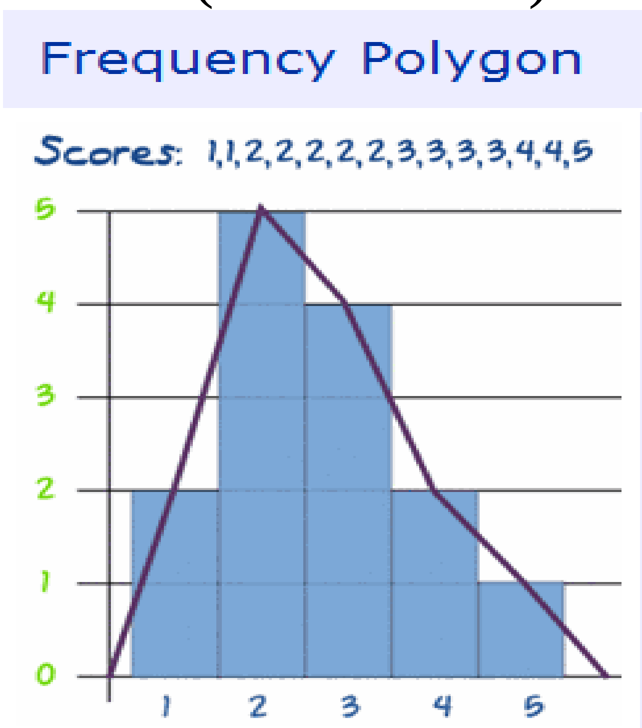
"The most common type is a graph in which the classes are plotted on the horizontal axis and the frequency of each class is plotted on the vertical axis. This type of graph is called a ___________or (loosely) a bar graph."

Example: Physician's Incomes (N=40)

Incomes ($'000)				
Class i	**Range**	**Frequency**	**Relative Frequency**	**Cumulative Frequency**
1	**80 ≤X < 100**		***0.050***	***2***
2	***100* ≤X < 120**		***0.150***	***8***
3	***120* ≤X < 140**		***0.200***	***16***
4	***140* ≤X < 160**		***0.150***	***22***
5	***160* ≤X < 180**		***0.075***	***25***
6	***180* ≤X < 200**		***0.325***	***38***
7	***200* ≤X < 220**		***0.050***	***40***

Two Graphs:

Frequency polygon: in addition to the histogram representation, a __________ polygon is constructed by drawing a straight line between the __________ of adjacent class intervals.
(Picture 1-21).

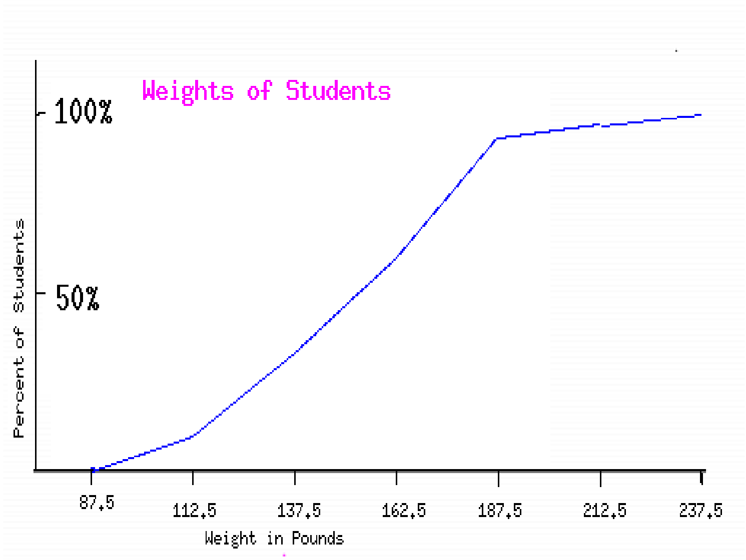


A graph made by joining the middle-top points of the columns of a frequency histogram

Ogive: with the __________ histogram, the ogive connects the ______ points.

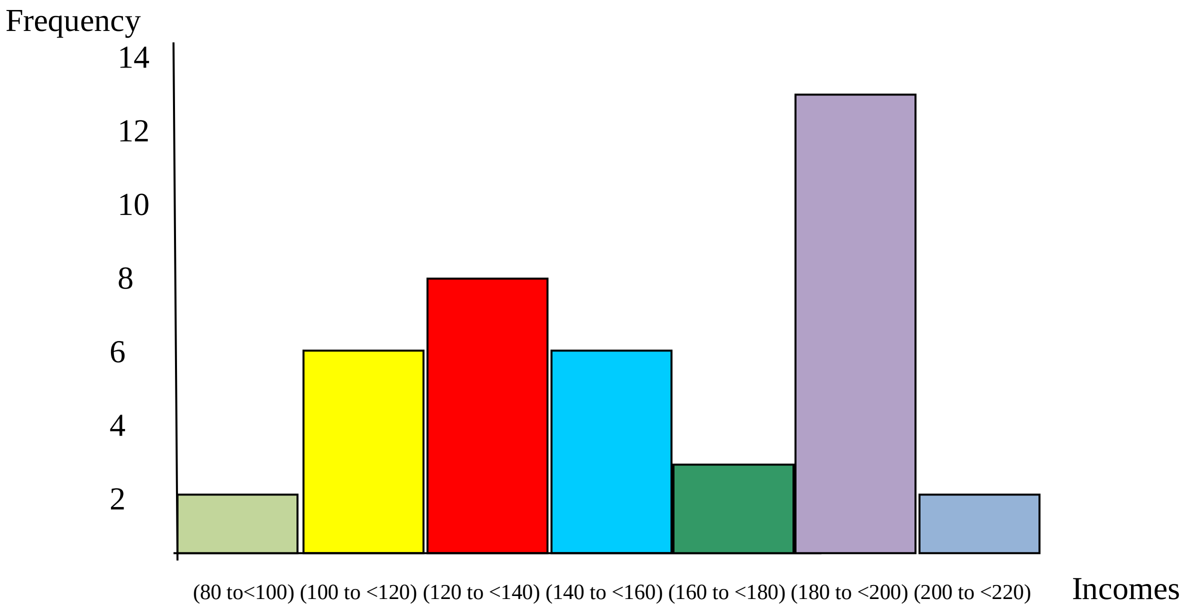
"Cumulative histogram can be "smoothed" by a line similar to the frequency polygon. This line is called a Ogive – connects the corner points of the cumulative histogram."

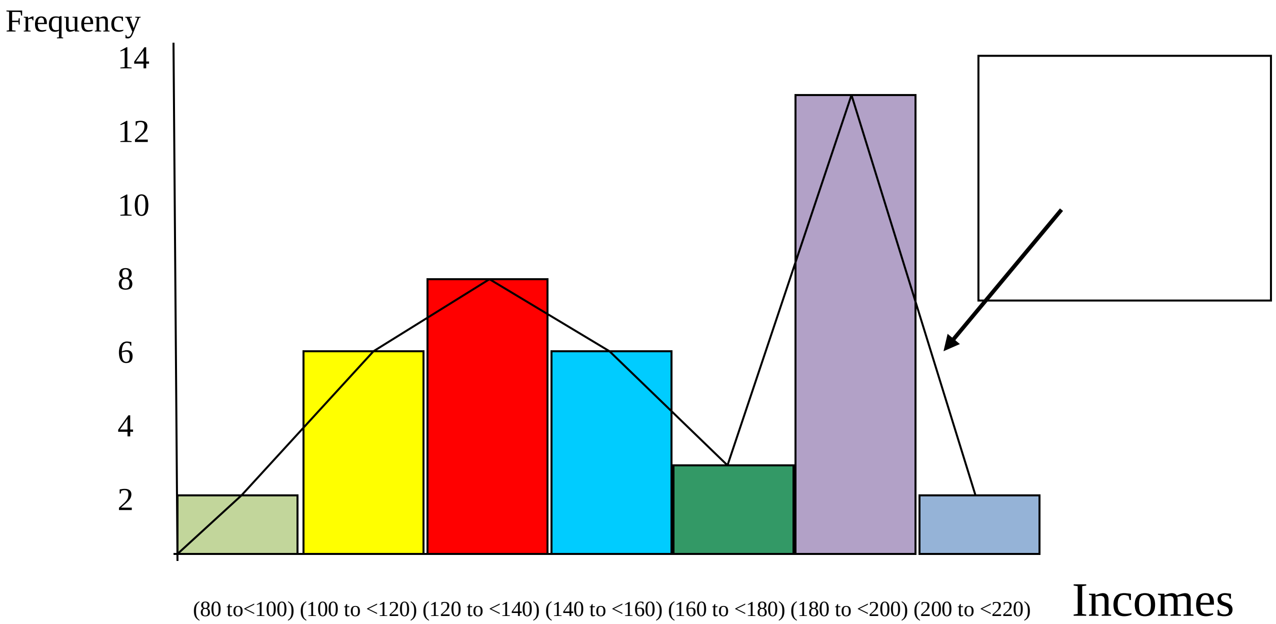
Note that the class boundaries rather than class marks are labelled, the cumulative number of individuals is read off the graph at the right boundary of the class, and straight (diagonal) lines are drawn accross each class. The information can also be displayed in a cumulative relative frequency ogive as indicated below.

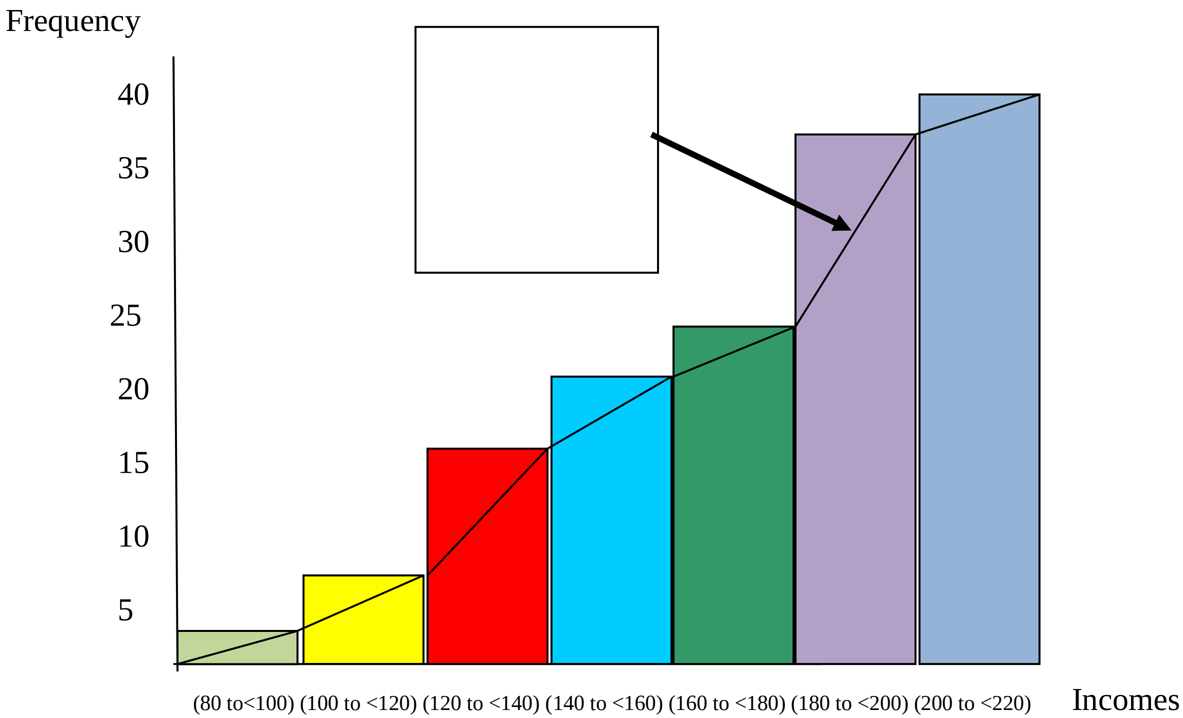


Example: Physician's Incomes (N=40)

Incomes ($'000)				
Class i	**Range**	**Frequency**	**Relative Frequency**	**Cumulative Frequency**
1	**80 ≤ X < 100**		***0.050***	***2***
2	***100*** **≤ X < 120**		***0.150***	***8***
3	***120*** **≤ X < 140**		***0.200***	***16***
4	***140*** **≤ X < 160**		***0.150***	***22***
5	***160*** **≤ X < 180**		***0.075***	***25***
6	***180*** **≤ X < 200**		***0.325***	***38***
7	***200*** **≤ X < 220**		***0.050***	***40***


Frequency
14
12
10
8
6
4
2
(80 to<100) (100 to <120) (120 to <140) (140 to <160) (160 to <180) (180 to <200) (200 to <220)
Incomes


Frequency
14
12
10
8
6
4
2
(80 to<100) (100 to <120) (120 to <140) (140 to <160) (160 to <180) (180 to <200) (200 to <220)
Incomes


Frequency
40
35
30
25
20
15
10
5
(80 to<100) (100 to <120) (120 to <140) (140 to <160) (160 to <180) (180 to <200) (200 to <220)
Incomes