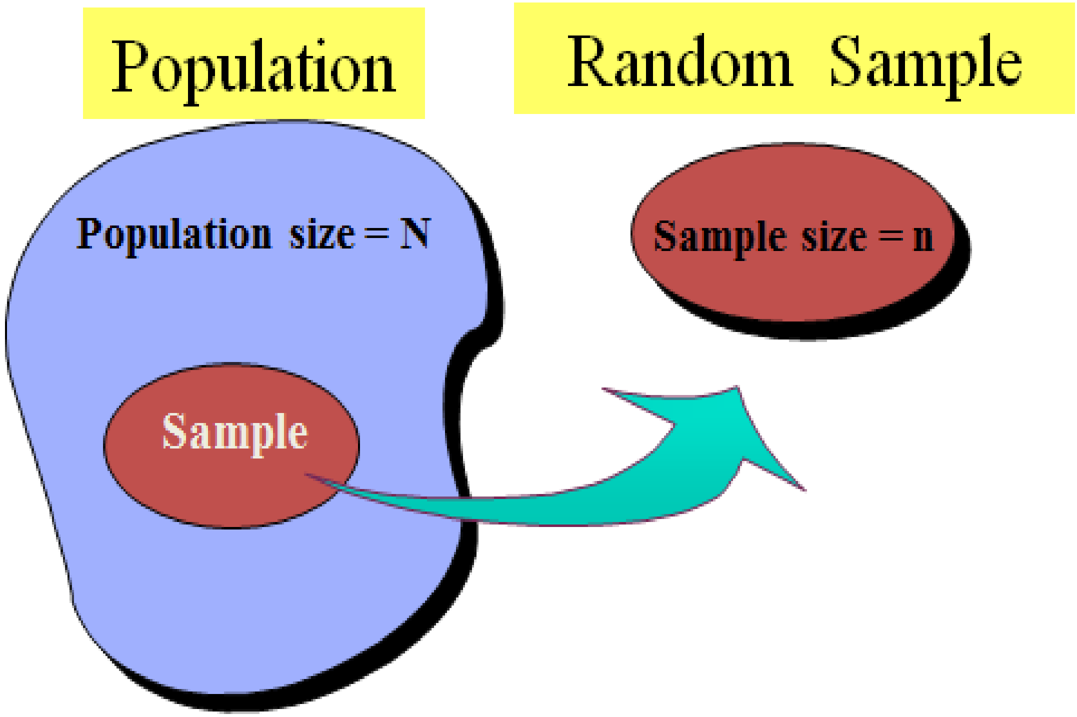
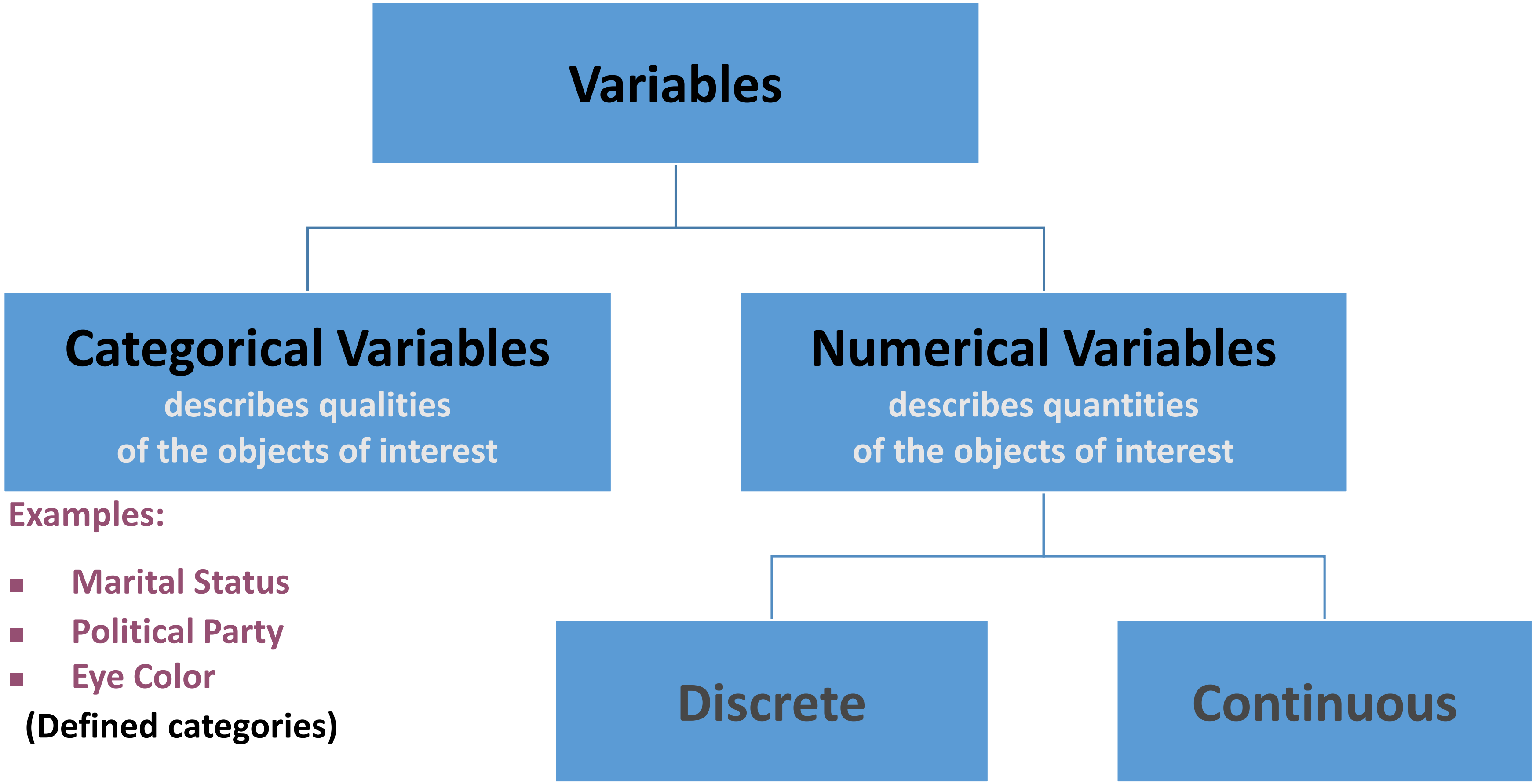
Summary---Topic 1: Introduction to Statistics



- The importance of sample:
- → Resource saving
- → Destruction of test units
- - Parameter: characteristics / measures describe population
- - Statistics: characteristics / measures describe sample

Types of Variables



Variables

Categorical Variables
describes qualities
of the objects of interest

Numerical Variables
describes quantities
of the objects of interest

Discrete

Continuous

Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

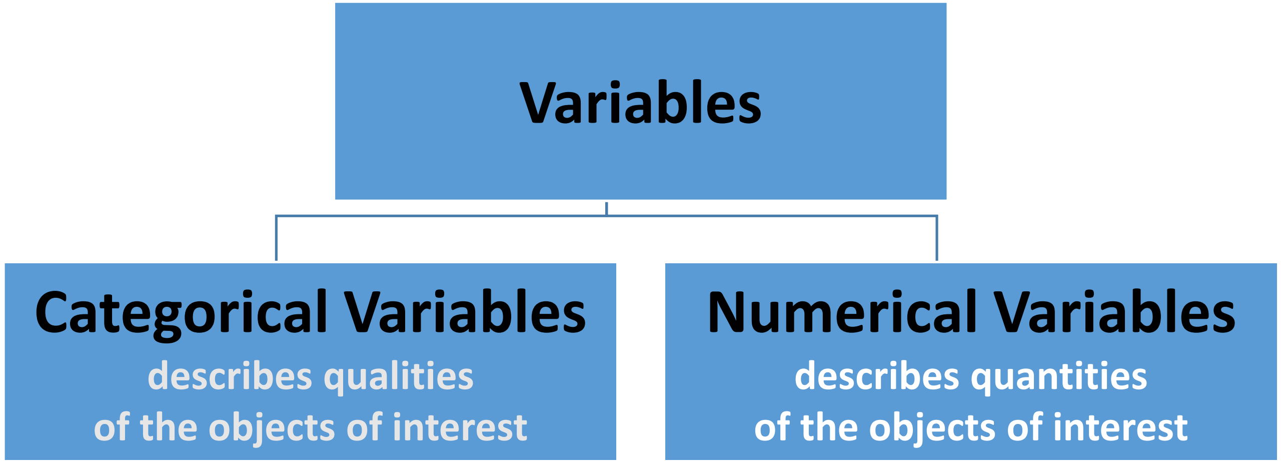
(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Organizing and Visualizing Data



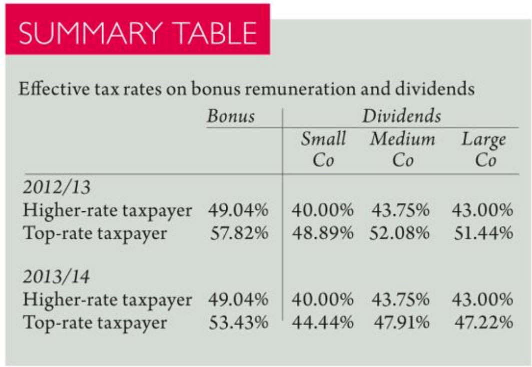
Variables

Categorical Variables
describes qualities
of the objects of interest

Numerical Variables
describes quantities
of the objects of interest

- Summary Table
- Bar Chart
- Pie Chart

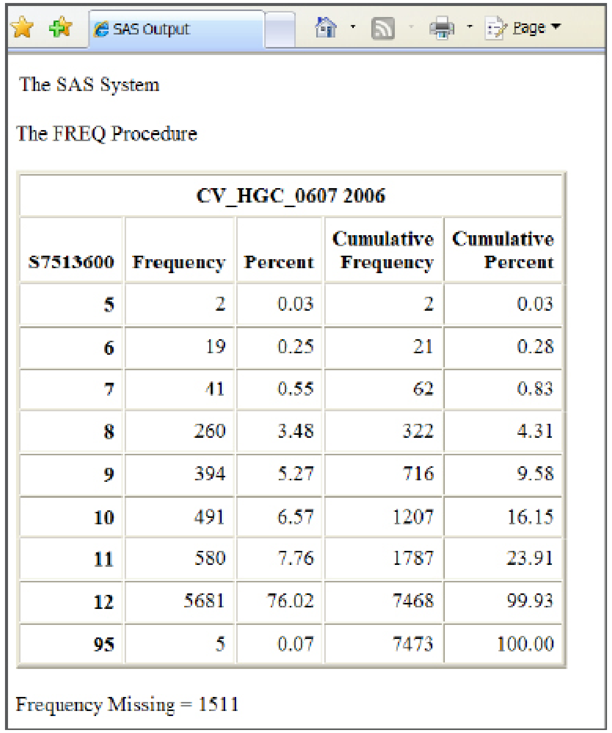
- Frequency Distribution
- Histogram



SUMMARY TABLE

Effective tax rates on bonus remuneration and dividends

	Bonus	*Dividends*		
		Small Co	*Medium Co*	*Large Co*
2012/13				
Higher-rate taxpayer	49.04%	40.00%	43.75%	43.00%
Top-rate taxpayer	57.82%	48.89%	52.08%	51.44%
2013/14				
Higher-rate taxpayer	49.04%	40.00%	43.75%	43.00%
Top-rate taxpayer	53.43%	44.44%	47.91%	47.22%



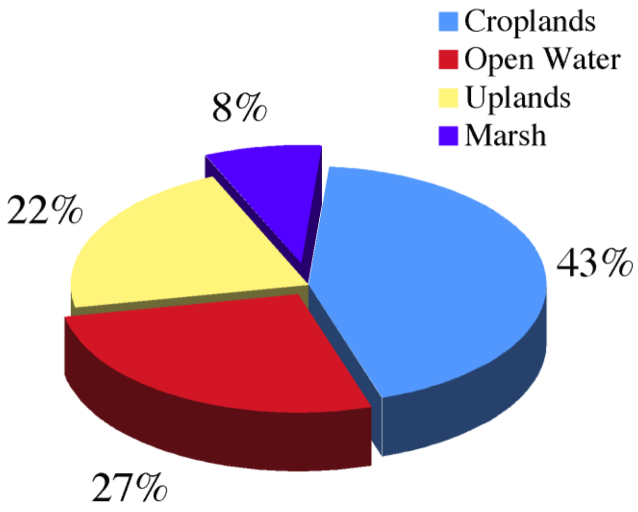
The SAS System

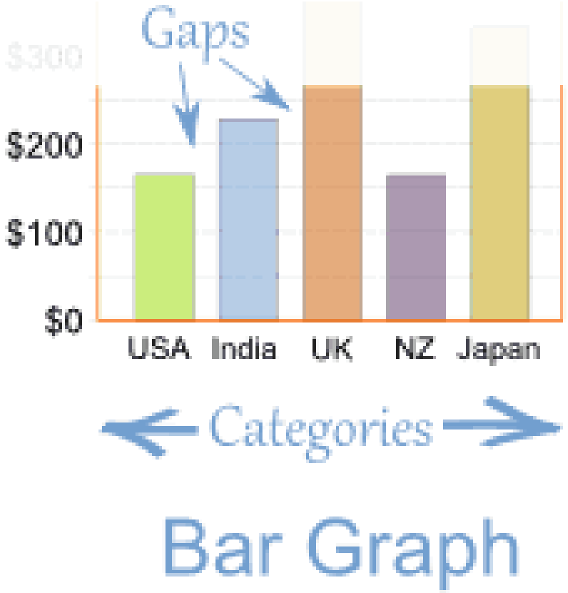
The FREQ Procedure

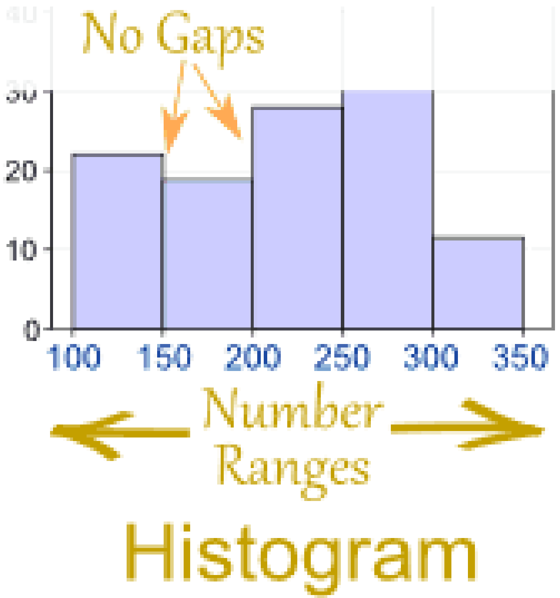
CV_HGC_0607 2006

S7513600	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5	2	0.03	2	0.03
6	19	0.25	21	0.28
7	41	0.55	62	0.83
8	260	3.48	322	4.31
9	394	5.27	716	9.58
10	491	6.57	1207	16.15
11	580	7.76	1787	23.91
12	5681	76.02	7468	99.93
95	5	0.07	7473	100.00

Frequency Missing = 1511







Principles of Excellent Graphs

1. The graph should not distort the data
2. The graph should not contain unnecessary adornments (chart junk)
3. The scale on the vertical axis should begin at zero
4. All axes should be labeled with proper scales
5. The graph should contain a title
6. The simplest possible graph should be used for a given set of data

Exercises and Solutions

Q1. For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous.

a) Number of cell phones in a household.

For example, 5 ➔ Numeric ➔ discrete

b) Length of the longest phone call made in a month

1.5 hours ➔ Numeric ➔ continuous

c) Whether the household has a land line.

Yes/No ➔ categorical

d) Whether there is a high-speed Internet connection in the household.

Yes/No ➔ categorical

Reminder*

Why are you in college? Answer:

1. Person Growth
2. Career Opportunities
3. Parental Pressure
4. Personal Networking

Results: 1, 4, 3, 2, 2, 1, 2, 3, 3, 1, 4, 2

Coding categorical data with numbers: Although the above data values are numbers, the variable is still categorical

Q2. The following data is about the cost of electricity (in \$) during July 2014 for a random sample of 50 one-bedroom apartments in a large city.

96	171	202	179	147	102	153	197	127	**82**
157	185	90	116	175	111	148	**213**	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

a) Construct a frequency distribution and a percentage distribution that have class intervals with the upper class boundaries \$99, \$119, and so on.

b) Construct a cumulative percentage distribution.

c) Construct a histogram.

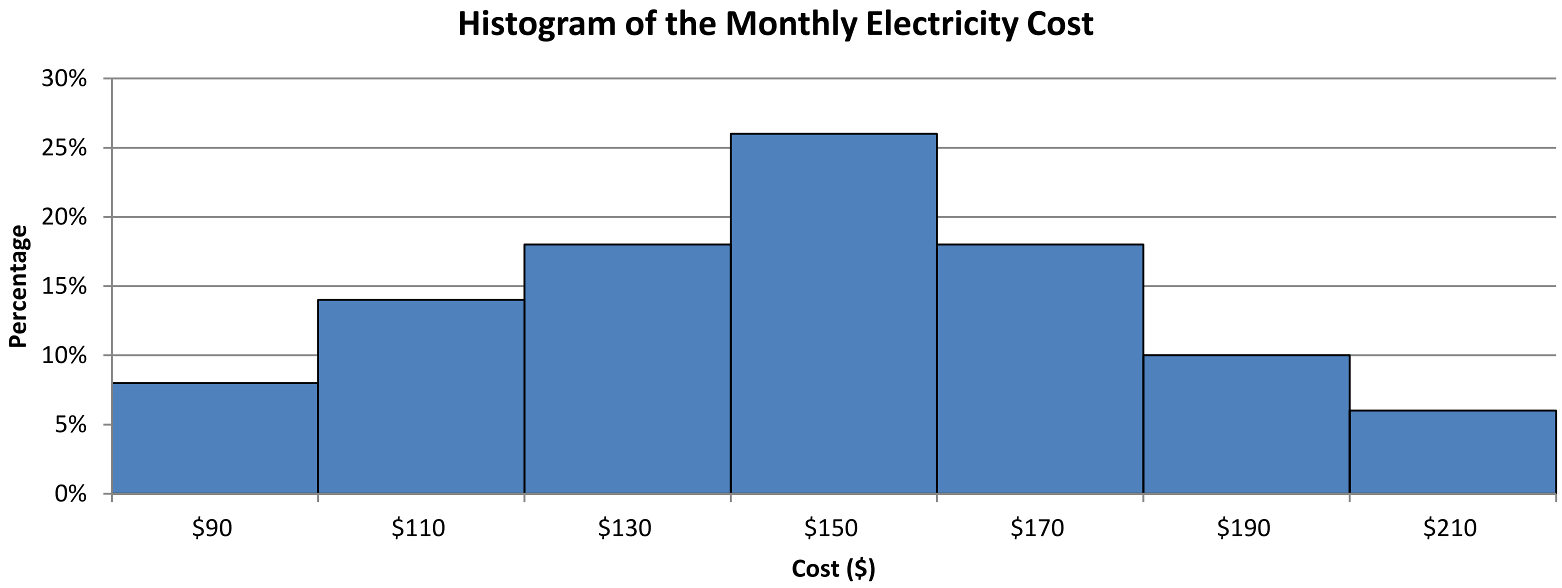
d) What is the total frequency of cost to be at least \$120 but less than \$180?

Steps to construct a frequency distribution:

1. Find the smallest and largest numbers in the data: 82, 213
2. Compute the range: 213 – 82 = 131
3. Determine the number of classes and the class interval (width):
   the class interval (width): 20 ➔ the number of classes =131/20 ≈ 7.
4. Determine class boundaries: $99, $119, $139, $159, $179, $199 $219,
5. Assign the observation to each class and count the number of observations

	Frequency	Percentage	Cumulative Percentage
$80 - Less than $100 (80-99)	4	4/50=8%	8%
$100 – Less than $120 (100-119)	7	7/50=14%	(8+14)%=22%
$120 – Less than $140 (120-139)	9	9/50=18%	(8+14+18)%=40%
$140 – Less than $160 (140-159)	13	13/50=26%	(8+14+18+26)%=66%
$160 – Less than $180 (160-179)	9	9/50=18%	(8+14+18+26+18)%=84%
$180 – Less than $200 (180-199)	5	5/50=10%	(8+14+18+26+18+10)%=94%
$200 – Less than $220 (200-219)	3	3/50=6%	(8+14+18+26+18+10+6)%=100%

c)



d) What is the total frequency of cost to be at least $120 but less than $180?

	Frequency	Percentage	Cumulative Percentage
$80 - Less than $100 (80-99)	4	4/50=8%	8%
$100 – Less than $120 (100-119)	7	7/50=14%	(8+14)%=22%
$120 – Less than $140 (120-139)	9	9/50=18%	(8+14+18)%=40%
$140 – Less than $160 (140-159)	13	13/50=26%	(8+14+18+26)%=66%
$160 – Less than $180 (160-179)	9	9/50=18%	(8+14+18+26+18)%=84%
$180 – Less than $200 (180-199)	5	5/50=10%	(8+14+18+26+18+10)%=94%
$200 – Less than $220 (200-219)	3	3/50=6%	(8+14+18+26+18+10+6)%=100%

9+13+9=31
observations
(or 31/50=62%).

Q3. Figure 1 below shows the profits of ABC company from 2000 to 2004. To show the company’s profit from 1990-2004 to shareholders, the managing director added the profit of the company in 1990 to the graph (Figure 2).

Do you think that the managing director is misleading the shareholders? Justify your answer.

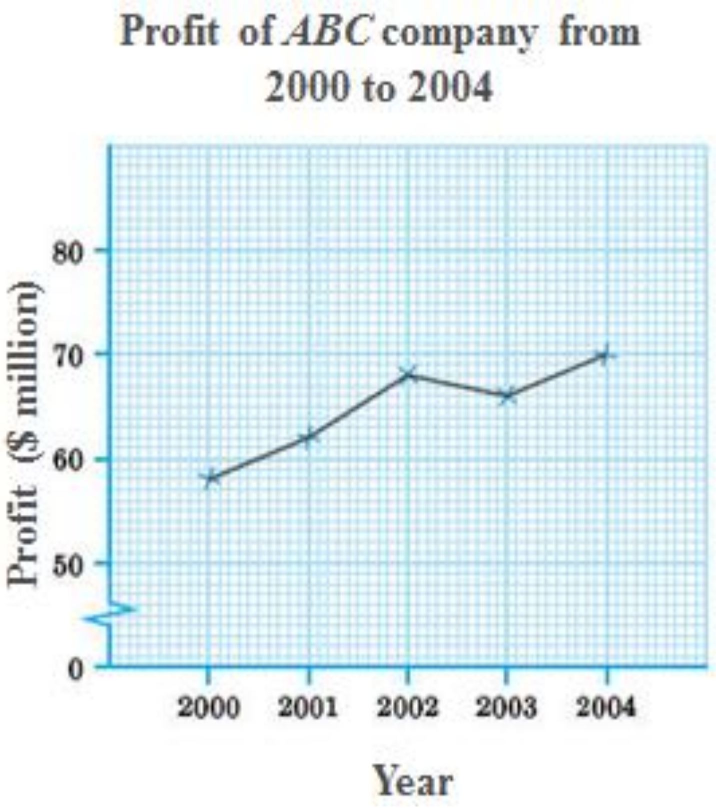


Fig1

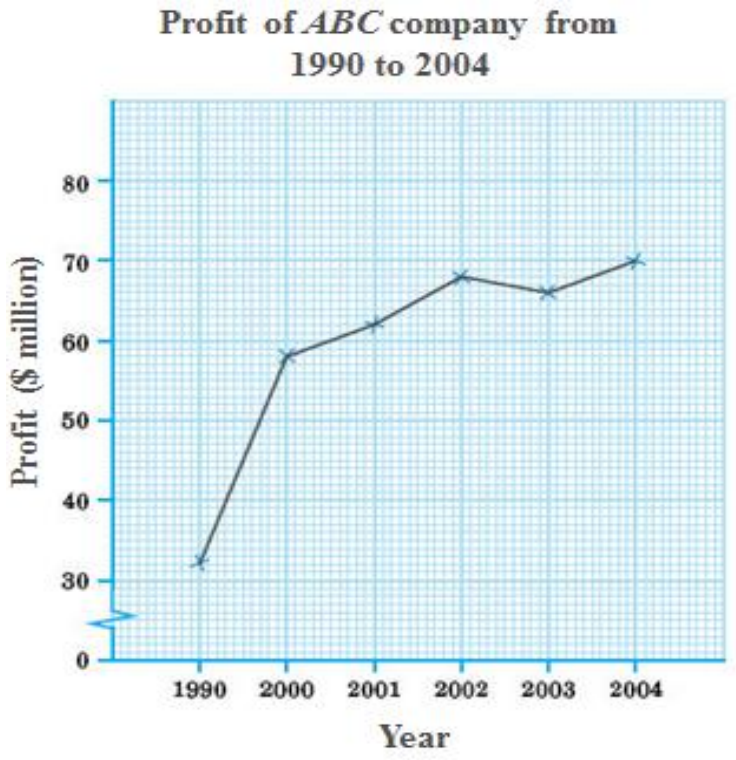


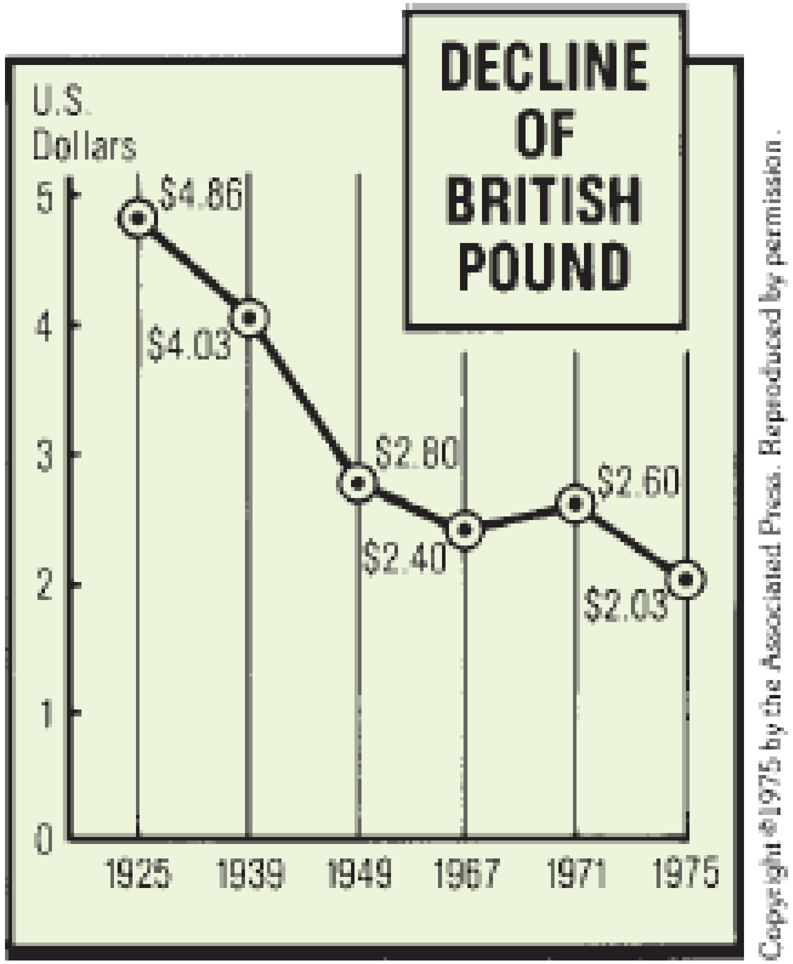
Fig2

Yes,

the managing director is misleading the shareholders in Figure 2 because **the profits in 1991 to 1999 are not shown.**

It gives the shareholder an impression that the company’s **profit increases rapidly from 1990 to 2000.**

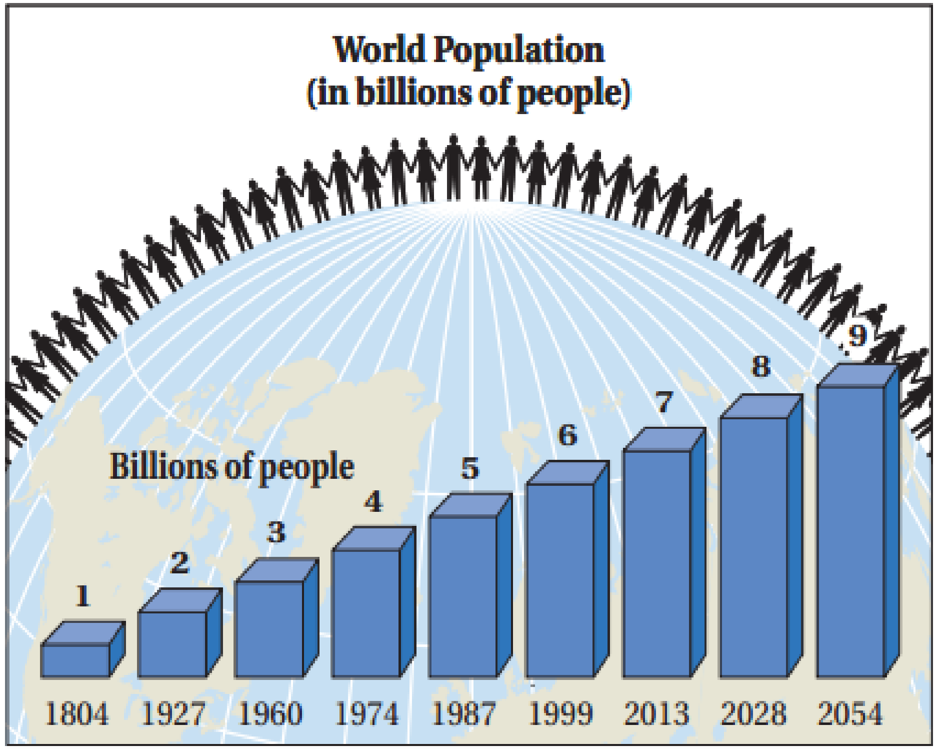
Q4. The graph below appeared in the Lexington Herald-Leader newspaper on 5th October, 1975.
Discuss the correctness of this graph.



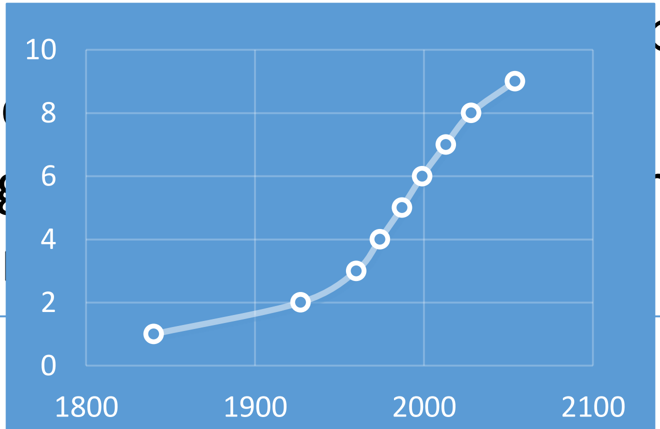
Copyright ©1975 by the Associated Press. Reproduced by permission.

- X-axis: The **intervals** of each two consecutive years are **not equal**. The bin widths range from 4 years to 18 years. It gives an impression that the British pound declined steadily from year 1971-1975 compared with year 1939-1949.
- The line and the data: The **measurement** unit is **not clearly stated**. Say in 1925, the $4.86, does it means US$4.86 to $1 British pounds.
- Title: too subjective.

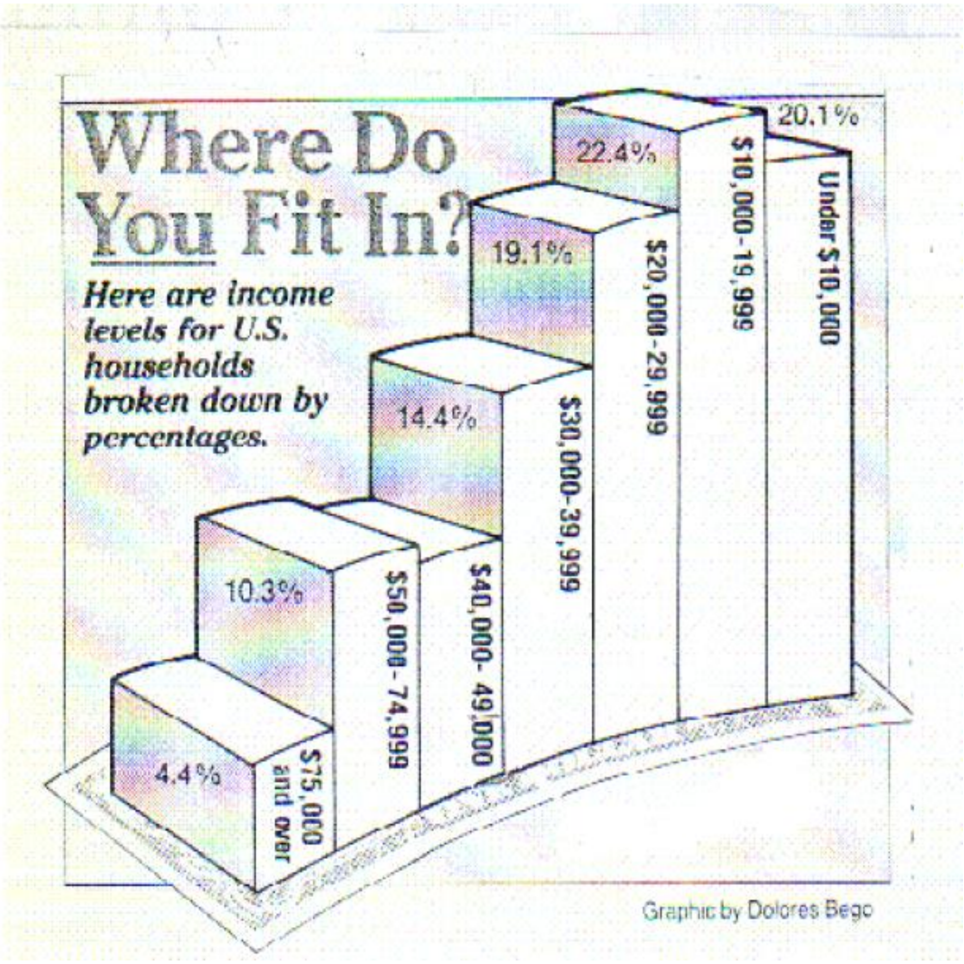
Q5. The following graph shows the world population from 1804-2054 (numbers for future years are based on United Nations projections).
Critique the graph in terms of its layout, content and clarity.



- The figures of people lining the globe do not give any information about world population. And it may give the impression that future world population will be declining. In the chart, **it appears that world population has been raising linearly**.
- Notice that the **time intervals** on the horizontal axis are **not uniform in size**.
- You shou ph showing
  slow incr at the
  beginning ncrease
  after year



Q6. The following graph shows the U.S. household income data in 1985. (Source: The U.S. Department of Labor).
Critique the graph in terms of its layout, content and clarity.



- The 3-D display makes it **difficult to read the bars**. Focusing at the front of each bar, the side of each bar, or the back of each bar will give different impressions.
- The x-axis goes from **right to left**, instead of the usual direction left to right, thereby giving a **misleading** perception of the asymmetry. Moreover, the curved and sloped x-axis **exaggerates the difference** between lower-income bars and upper-income bars.
- **Percentage figure was missed** from the bar of $40,000-$49,000. Moreover, the upper boundary of this bar should be $49,999.
- The bar widths are not proportional to the interval ranges. For example, it goes by $10,000 then by $25,000, increasing the height of the $50,000-$74,999 bar.
- Total number of households was not given.