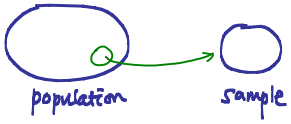
September 9 2010, Lecture 1:

CHAPTER 1: DATA AND DISTRIBUTIONS

Populations, Samples, and Processes (Section 1.1, page 3)

Question: What is a statistic? A numerical (description) summary measure based on data from a sample

The basic Idea behind all statistical methods of data analysis is to make inferences about a population by studying a relatively small sample chosen from it.



STATISTICS deals with the collection, description & interpretation of DATA.

Data may arise naturally as observations of "things as they are", or as a result of a controlled designed experiment.

As mentioned, statistical methods are based on the idea of analyzing a **sample** drawn from a **population**. For this to work, the collection of data must be done carefully and according to certain rules. The important thing to keep in mind is that data collected incorrectly may invalidate any conclusion we may draw from it.

Definition: a ***population*** is the set of all measurements or objects of interest in a particular study.

Example: I want to test the average alcohol consumption in all adults (19 and over) residing in Oshawa. What would my population be?

all adults (19+) residing in Oshawa

Definition: a ***sample*** is a subset of the population.

Definition: a ***simple random sample*** (srs) is a sample chosen in such a way that each member of the population has an equal chance of being chosen.

Reason for using a srs? more accurate representation of the population
⇒ more accurate statistic.

For example: Let's say we wish to test the heights of students at UOIT by measuring a sample of 100 students. How should we choose the 100 students to measure? Some methods are obviously bad:

choosing the sample which represents only a certain type of population (basketball, Volleyball players)

Thus, a **simple random sample** is ideal.

Question: If the population is the thing that interests an investigator, why doesn't he/she just simply examine each member in the population?

too much time !! (All 19+ adults in oshawa ?)
too many restrictions (locations . . .)

Thus to obtain information about the population the investigator must use a **simple random sample**.

Example: A candidate for political office wishes to assess his chances in the next election by estimating the proportion of voters in his district who will vote for him. He obtained a srs of 1000 voters and found that 600 voters in the sample intended to vote for him and 400 against him.

What is the population? All voters in the district.

What is the sample? 1000 voters (srs)

Estimate the proportion of voters in the population that intend to vote for this candidate. Would you expect your estimate to be accurate? Explain.

- Probability (vote for him) $= \frac{600}{1000} = 0.6 = 60\%$

- It is accurate since the srs is quite large enough.

TYPES OF DATA

In statistics one works with several kinds of data, so it's important to differentiate between these types:

Univariate data: consists of observations on a single variable.

lifetime of an automobile tire

Bivariate data: when observations are made on each of two variables.

measuring height & weight for each basketball player on a team

Multivariate data: when observations are made on more than two variables.

measuring height, weight, wingspan for each basketball player on a team

NUMERICAL (QUANTITATIVE) DATA is data measured on a numerical scale. It is used to tell us how much or how many. There are two types of numerical data:

> **Continuous data** is data such as length or weight, which in theory, can be measured to any degree of accuracy.

Discrete data is data whose possible values are isolated points on the number line. For example, data such as the number of children in families or number of fish caught by fisherman is discrete data.

CATEGORICAL (QUALITATIVE) DATA is data that records qualities of persons or objects, such as names, gender, or student numbers. There are two types of categorical data:

Ordinal data is data which has a natural ordering.

e.g) grade (A, B, C, D, F),

Non-ordinal data is data which has no natural ordering.

name of car brand.

Example: State what type of data would be recorded in each of the following cases.

(a) In a sample of 1000 Canadian voters each is classified as being Liberal, Reform, NDP, or other.

Categorical and Non ordinal

Numerical < Continuous, Discrete

(b) A biologist measures the height of tomato plants.

Numerical and Continuous

Categorical < ordinal, non ordinal

(c) In the automobile industry, quality control classifies each paint job as excellent, good, fair or poor.

Categorical and ordinal

One more definition: A ***variable*** is a characteristic of a person or a thing that can be assigned a number or a category.

Age or height

Let x_1 = Age, x_2 = height . .

Example: For each of the following settings (i) identify the variable(s) in the study, (ii) explain the type of data for each variable in the study (e.g., categorical and ordinal, continuous or discrete, etc.), (iii) identify the observational unit, and (iv) determine the sample size.

(a) A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*.

i) Width (molar)

ii) univariate, Numerical, continuous

iii) a molar

iv) $n = 36$

(b) The birthweight, date of birth, and the mother's race were recorded for each of 65 babies.

i) birthweight, DoB, Mother's race

ii) multivariate, BW: Numerical, continuous
DoB: " , discrete (or Categorical, ordinal)
M.R: Categorical, non ordinal

iii) a baby + mother

iv) $n = 65$

Once your data is collected, it is often useful to describe the data through displays or summaries that bring attention to certain characteristics which may be of interest to us.

Visual Displays for Univariate Data (Section 1.2, page 8)

Frequency and Relative Frequency Distributions:

A set of raw data gives us very little information as to how the observations are distributed. For example, consider the following sample of grades from a large mathematics class (the grades have been ordered from lowest to highest for convenience).

55	55	55	56	56	57	57	57	58	59	59	60	60	60	61	65
65	66	66	66	67	67	67	67	67	67	68	68	68	69	69	69
69	69	69	70	70	70	71	71	72	72	72	72	73	73	73	73
73	73	73	74	74	74	74	76	76	76	76	76	76	76	77	77
77	77	78	78	79	79	79	80	80	80	80	81	82	82	82	83
83	83	83	84	84	85	85	87	87	88	88	89	92	92	94	

What is the distribution of these grades? Are there more B's than C's? Do the grades clump around a certain grade?

$50 \le x < 60$

50 60

Divide the sample of grades into five classes, say , [50, 60), [60, 70), [70, 80), [80, 90), and [90, 100), and find the frequency and relative frequency for each class.

Frequency and Relative Frequency Distribution (or Table)

Class Interval	Class Frequency (f_i)	Class Relative Frequency (f_i/n)
[50, 60)	11	11/95 = .1158
[60, 70)	24	24/95 = .2524
[70, 80)	36	36/95 = .3789
[80, 90)	21	21/95 = .2211
[90, 100)	3	3/95 = .0316

# of sample values in each class

proportion of sample values in each class

Lower Class Boundary

Upper Class Boundary

total (n) = 95 total # of sample

total = 1 (95/95 = 100%)

Note: The **CLASS MID-POINT** is the average of the lower and upper class boundaries.

e.g) for first class [50, 60), the class mid-point is

$\frac{50+60}{2} = 55$ ~ representing a typical value in the class

Note: The **CLASS WIDTH (CW)** is usually (but not always) the same size within a histogram.

CW = UCB − LCB , for the first class [50, 60)

CW = 60 − 50 = 10

→ for the example above, all classes have CW = 10.

Frequency and Relative Frequency Histograms

A frequency (or relative frequency) histogram of a sample is simply a picture of the frequency (or relative frequency) distribution of that sample.

Example (FREQUENCY HISTOGRAM): To draw a frequency histogram of our sample of 95 grades from a large mathematics class we proceed as follows:

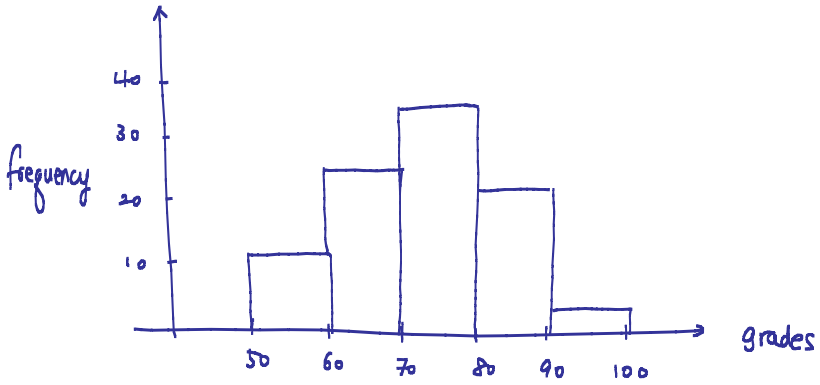
(a) Form an (x,y) coordinate system with the class intervals on the horizontal axis (x) and the class frequencies marked on the vertical axis (y).

(b) Above each class draw a rectangle, whose height is equal to the class frequency.

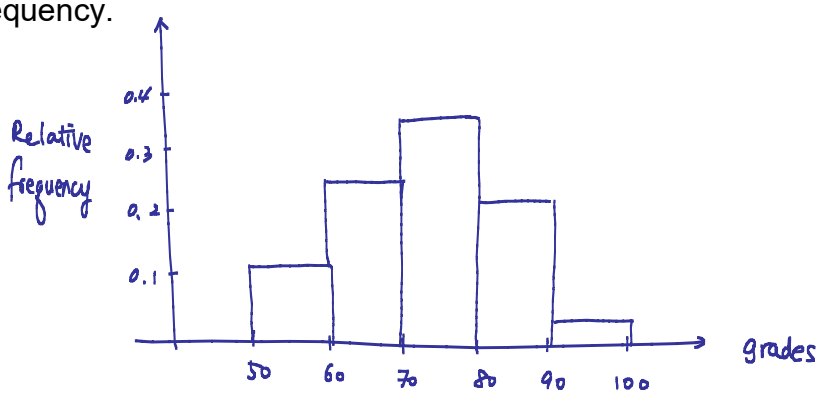
Recall:

x y

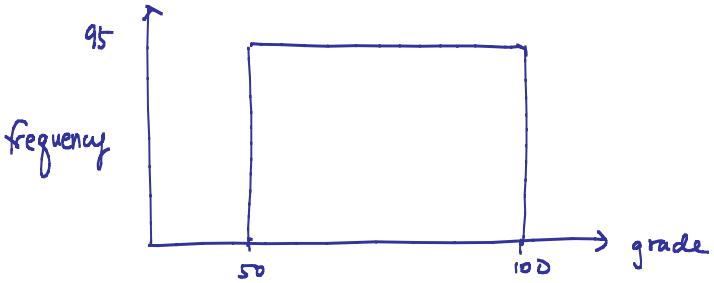
Class Interval	Class Midpoint (m_i)	Class Frequency (f_i)	Class Relative Frequency (f_i/n)
[50, 60)	55	11	11/95 = .1158
[60, 70)	65	24	24/95 = .2526
[70, 80)	75	36	36/95 = .3789
[80, 90)	85	21	21/95 = .2211
[90, 100)	95	3	3/95 = .0316



To draw a relative frequency histogram, mark the relative frequencies on the vertical axis and above each class draw a rectangle whose height is equal to the class relative frequency.



Note: The class width determines the number of classes. This choice will affect the "look" of the histogram. For example, what would the frequency histogram of the class grades look like if we used one class [50, 100).



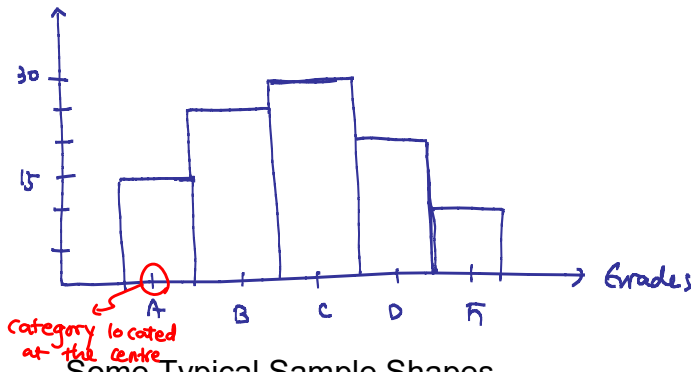
not meaningful

Histograms for Categorical Data

For a categorical data, the categories themselves may be used as classes.

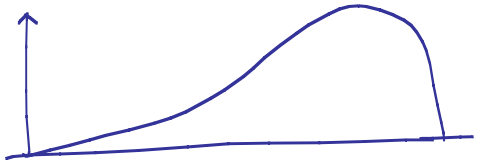
Example: The table below gives the final grades to a class of 100 students in an elementary statistics course. Create a categorical histogram.

Grades	A	B	C	D	F
# of students	15	25	30	20	10

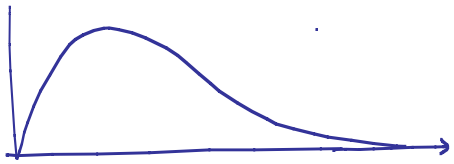


Some Typical Sample Shapes

(1) A **LEFT SKEWED** sample is one whose histogram (or stem and leaf plot) has a long left tail; the sample values tend to cluster at the right end of the scale and taper off at the lower end.



(2) A **RIGHT SKEWED** sample is one whose histogram (or stem and leaf plot) has a long right tail; the sample values tend to cluster at the left end of the scale and taper off at the higher end.



(3) A **SYMMETRIC** sample is one whose histogram (or stem and leaf plot) is distributed approximately the same on each side of some central value.

(4) A particular type of symmetric sample is a **BELL- SHAPED**; all bell shaped samples are symmetric, but not all symmetric samples are bell shaped.