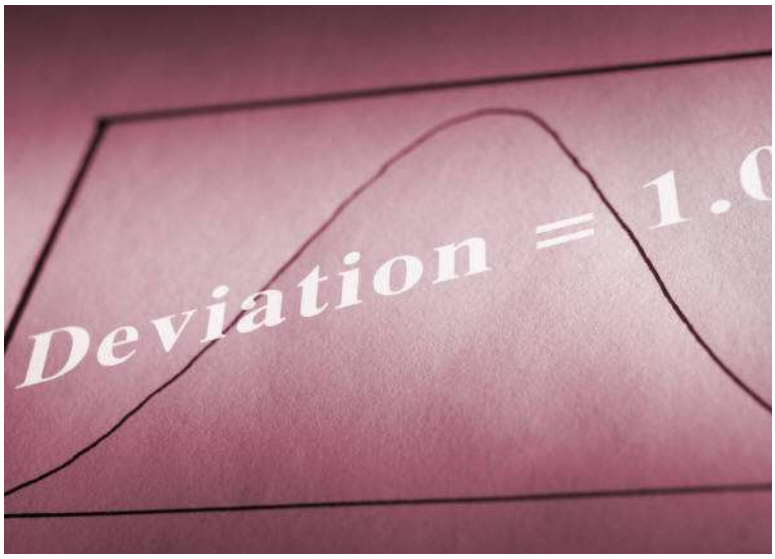


OPIM303 - Managerial Statistics

Descriptive Statistics: Collection, summarization, and characterization of a data set.

Inferential Statistics: estimates a characteristic of a set of data. Helps uncover patterns in data sets which are unlikely to occur by chance.

Samples: the portion of the population which is selected for analysis.

Population or Universe: the totality of items or things under consideration.

Statistic: a summary measure computed from sample data that is used to describe or estimate a characteristic of the entire population.

Parameters: a summary measure that describes a characteristic of an entire population.

Primary Source: the data collector is the one using the data for analysis.

Secondary Source: one organization or individual has compiled the data for use by another organization or individual.

Frame: a complete or partial listing of items comprising the population.

Focus Group: a marketing tool used to solicit unstructured responses to open ended questions.

Nonprobability Sample: the items or individuals included are chosen without regard to their probability of occurrence.

Probability Sample: the subjects of the sample are chosen on the basis of known probabilities.

Sampling with Replacement: a selected item is returned to the frame where it has the same probability of being selected again.

Sampling without Replacement: an item, once selected, is not returned to the frame and therefore cannot be selected again.

Systematic Sample: the N items in the frame are partitioned into k groups through division by the sample size n. $k = N/n$

Cluster Sample: N items in a frame are divided into several clusters so that each cluster is representative of the entire population. Natural examples include countries, election districts, city blocks, apartment buildings, and families.

Nominal Scale: classifies data into various distinct categories in which no ordering is implied.

Ordinal Scale: classifies data into distinct categories in which the ordering is implied.

Interval Scale: an ordered scale in which the difference between two measurements is a meaningful quantity which does not involve a true zero point.

Ratio Scale: an ordered scale in which the difference between the measurements involves a true zero point such as height, weight, age.

Coverage Error: occurs if certain groups of subjects are excluded from the frame listing so that they have no chance of being selected in the sample. Results in ***selection bias***.

Nonresponse Error: arises from the failure to collect data on all subjects in the sample. Results in ***nonresponse bias***.

Educational Goals for the Course

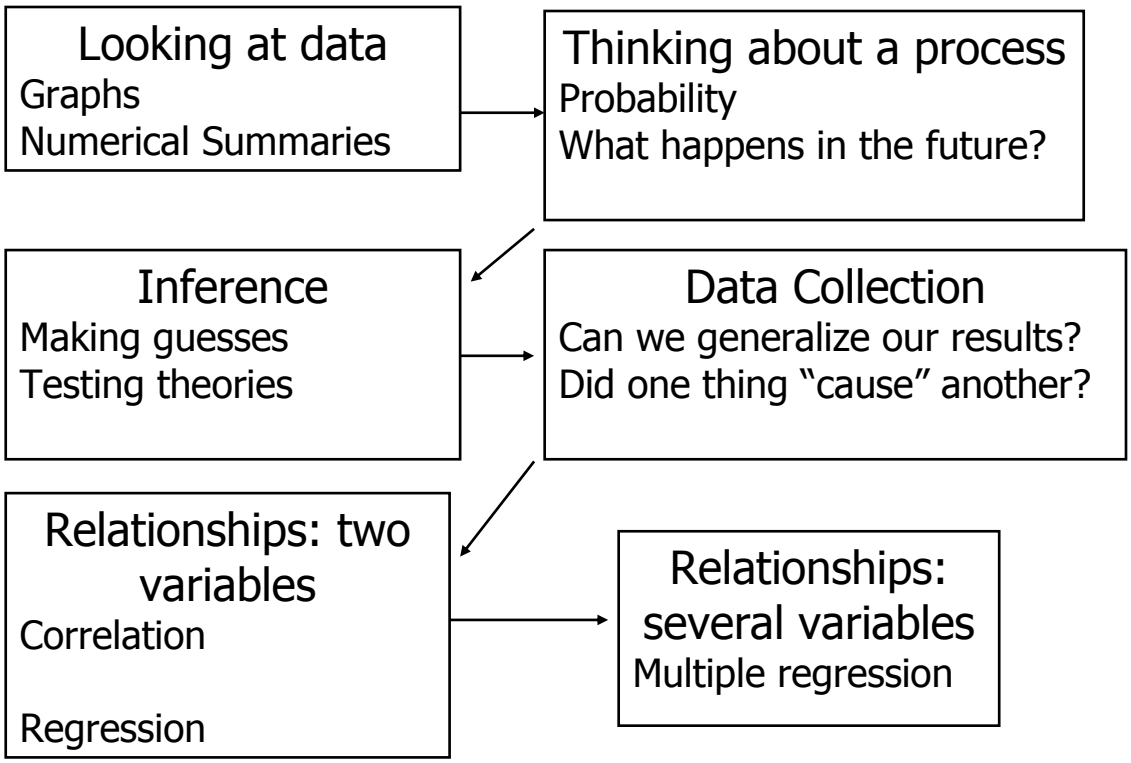
- "Poets"
  - Learn some basic principles of data analysis.
  - Learn enough so that you will know what your
- "Engineers"
  - Sharpen quantitative skills.
  - Practice explaining technical material by helping non-technical

Both groups should get something out of this class.

Technical Goal

- Understand Multiple Regression
  - Relationships between variables.
    - Is weight or length more important for determining a car's gas mileage?
  - Predict one variable using others.
    - What profit should you expect when building a house with X sq. ft. and Y bedrooms.
  - Evaluate performance in the face of mitigating circumstances.
    - Set production goals for subordinates with different staff sizes, regional economic conditions, etc.

From here to there…



Class Resources

- Textbook
  - Does a better job explaining theory.
  - Not as many examples.
- PowerPoint presentation
  - Usually available before our meetings
- Lectures
  - My chance to explain what the reading is supposed to say.
- Homeworks/Examples

Intro to MS Excel

- I assume you know how to:
  - Create formulas in cells
  - Copy, cut and paste cells within a workbook
  - Make simple charts
    - E.g., Pie charts, bar charts
  - Have the Data Analysis Add-in installed
  - Do basic formatting
    - E.g., numbers, percentages, etc.

Statistical Methods

- Descriptive statistics
  - Collecting and describing data
- Inferential statistics
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics: the type we hear about all the time in the media. Housing prices, unemployment, just basically numbers.

We will start by collecting data and describe the data using descriptive statistics. Averages for instance.

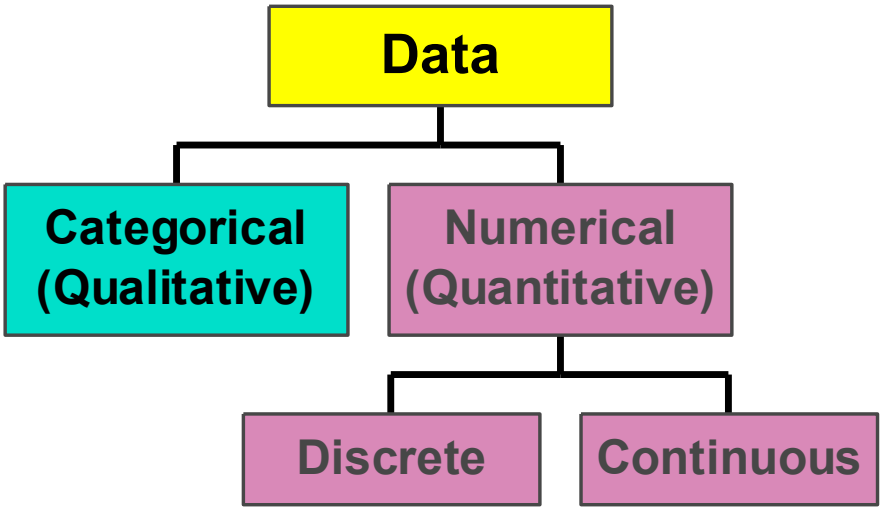
But the bulk of the course has to do with **Inferential Statistics** which has to do with drawing conclusions and making decisions about a population based only on SAMPLE DATA. Example, we have a list of the sales price of the houses in our area for a quarter. How can we use this data to infer the value of our house or any other house in the area? This is Inferential Statistics, we go one step beyond what the data is giving us.

This applies to business. For instance marketing and trying to decide if a new product will be a success. Collect data and conduct test on a sampling of consumers. But you want to know what the interest will be of all consumers. This is where inferential statistics comes in, going beyond the data we have.

Inferential Statistics break down into two parts: **Estimation** & **Hypothesis Testing**. Both are very closely related. In **Estimation** we want to estimate some number for the whole population given some sample data.

In estimation we want to come up with a number which accurately represents the whole population based upon a sample of the population. Hypothesis Testing is closely related to estimation. In Hypothesis Testing somebody gives you a number and you must prove or disprove it. Say it is proposed, based on estimation, that 80% of people prefer a certain new soft drink. We want to prove or disprove that the 80% number is accurate. We do this through Hypothesis Testing.

Types of Data



Simple Test:
Does it make sense to do arithmetic on it? If yes, it's numeric.

When considering if a numerical data type is Continuous consider the nature of the data. Time is always a continuous data type even if it is subdivided into age rounded to the year.

First we collect data then we examine it. May times we "assume" a certain set of data. In these examples we use the House data (actual data from early 1980's).

house	sqft	beds	baths	heat	style	garage	basement	age	fire	price	school
1	1,238	3	2	0	0	1	1	12	1	59.900	1
2	1,707	3	2	1	0	2	0	13	1	64.000	0
3	1,296	4	2	0	0	2	1	17	0	66.500	0
4	1,320	3	2	0	0	2	1	11	1	66.500	0
5	1,210	3	2	0	0	1	0	6	1	66.900	0
6	1,296	3	2	0	0	2	1	17	1	68.000	0
7	1,765	3	2	0	0	2	1	20	0	68.500	0
8	1,725	4	3	0	0	2	1	12	0	69.000	0
9	1,794	4	2	0	0	2	1	18	0	70.950	0
10	1,294	3	2	0	0	2	0	13	1	71.000	0

When we look at data we can distinguish between two broad categories. These are **Categorical Data** and the other is **Numerical Data**.

Numerical Data consist of numbers (weight, height, salary). Categorical data tends to be grouped by an attribute of an item (male/female [gender], red/blue[color], circle/square).

How can we decide if data is numerical or categorical? One way is to ask yourself, "does it make sense to do arithmetic on it?" For example, we can add two salaries and the result has a clear meaning, so that is numerical data. But it makes no sense to add two phone numbers, the result would have no meaning. So phone numbers are categorical data. Same for social security numbers, categorical. Many time categorical data is expressed as numbers. There is no average zip code or social security number, these must be categorical.

Now we can distinguish between two types of numerical data: Discrete and Continuous.

Discrete: can only take on discrete values. Examples include number of children in a household. This must be a whole number, a discrete value. Number of cars in a household as well.

Continuous: can take on any value. A persons weight for example.

Lets classify the headings of the house data spreadsheet according to data type.

house	sqft	beds	baths	heat	style	garage	basement	age	fire	price	school
1	1,238	3	2	0	0	1	1	12	1	59.900	1
2	1,707	3	2	1	0	2	0	13	1	64.000	0
3	1,296	4	2	0	0	2	1	17	0	66.500	0
Index Number	Numerical & Continuous	discrete	discrete	Categorical (coded)	categorical (coded)	Discrete Numerical	Categorical	Continuous	Categorical	Continuous	Categorical

Sqft: is rounded but could be taken to decimal places.
Bathrooms: real estate agents will apply ".5" but really discrete.
Heat: the number represents a category.
Garage: the number of cars it can hold.
Basement: presents or absence.
Age: time is continuous even if they round to the closest year.
Fire: presence or absence of a fireplace.
House: this is actually **categorical** even though it is an index number. It is basically a label which happens to be a consecutive number. Arithmetic on these numbers would have no meaning.

With Continuous variables you will always have to restrict yourself to a certain approximation, a certain precision you will round to. But they are still continuous.

Why do we distinguish between numerical data and categorical data? We do this because given a data set we want to examine it. Find it's average, it's common values, its max and min, etc. Weather the data is numerical or categorical will determine the type of tools we use on it.

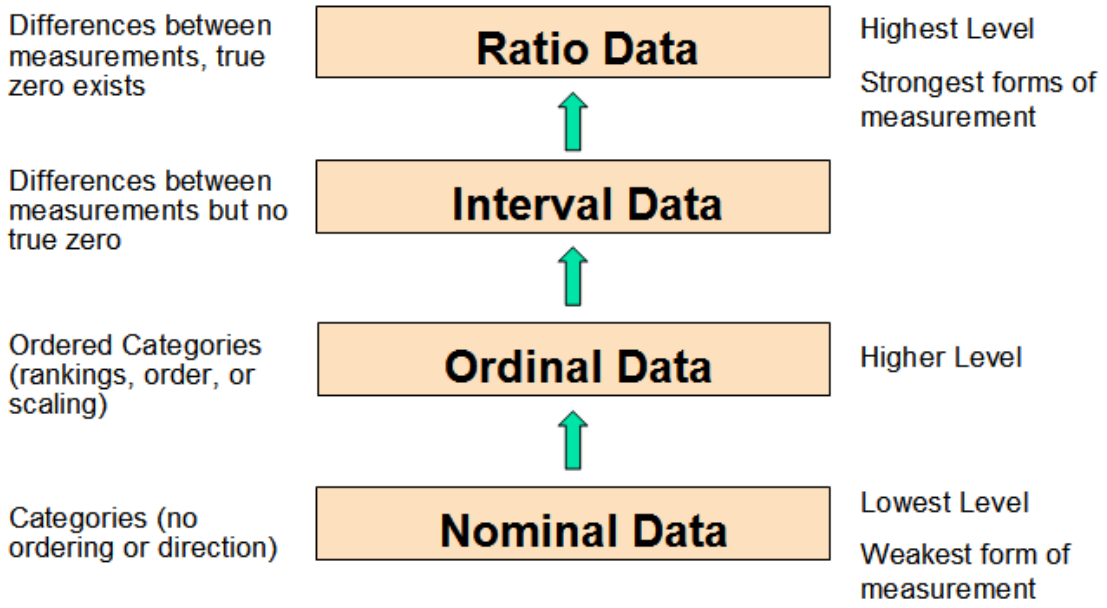
Scales

Nominal Scale: classifies data into various distinct categories but no order or ranking is implied. Examples include Yes/No, Democrat/Republican.

Ordinal Scale: classifies data into distinct categories in which ordering is implied. Examples include student grades, product satisfaction. Still, the ordering only implies which category is "greater," "better," or "more preferred". Not a hard measure.

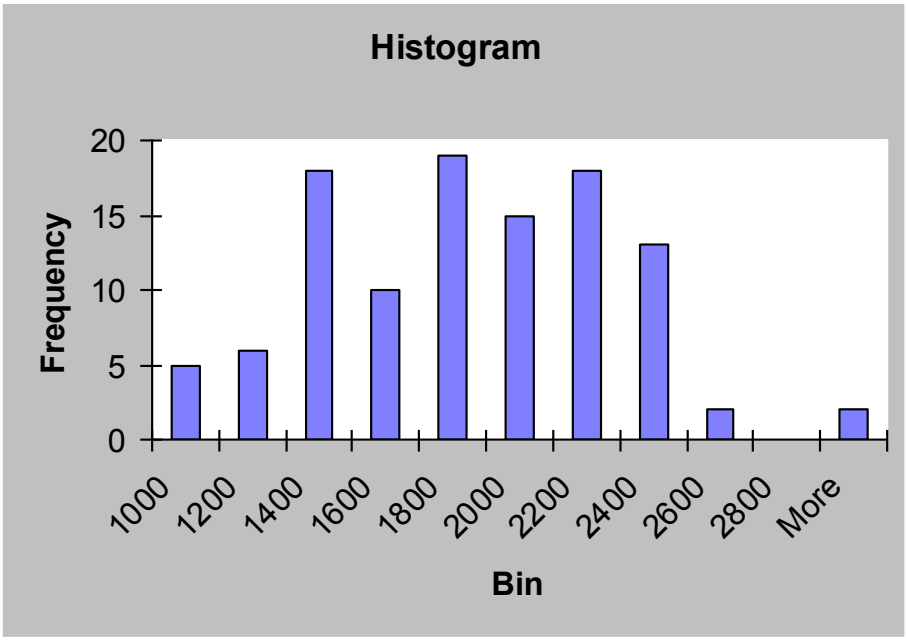
Interval Scale: an ordered scale in which the difference between two measurements is a meaningful quantity which does not involve a true zero point.

Ratio Scale: an ordered scale in which the difference between the measurements involves a true zero point such as height, weight, age.
Numeric data (interval or ratio) is a stronger measure than ordinal data.



Displaying Numerical Data

- Histogram example



The notes online include an Excel tutorial on histograms in PDF format.

Histogram: good for displaying numerical data. Notice the axis's. Frequency is how often a particular x category or range occurs (always on the vertical axis). Bins are the ranges which the x axis is broken into. For instance, 5 houses with up to 1200 square feet and 6 houses with between 1200 and 1400 square feet. No houses with between 2800 and 3000 (More) square feet and about 2 with more than 3000 (more) square feet.

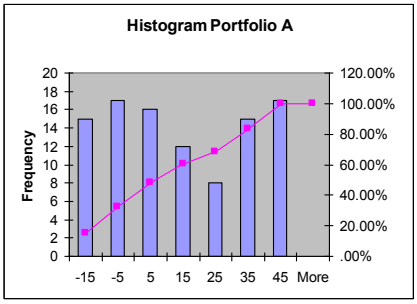
In this case we see that the typical house has between 1800 and 2000 square feet (19). This is the mean.

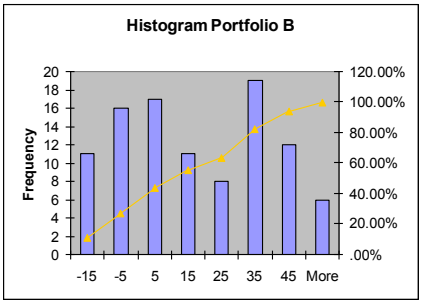
Sometimes the vertical axis is given in percent of observations.

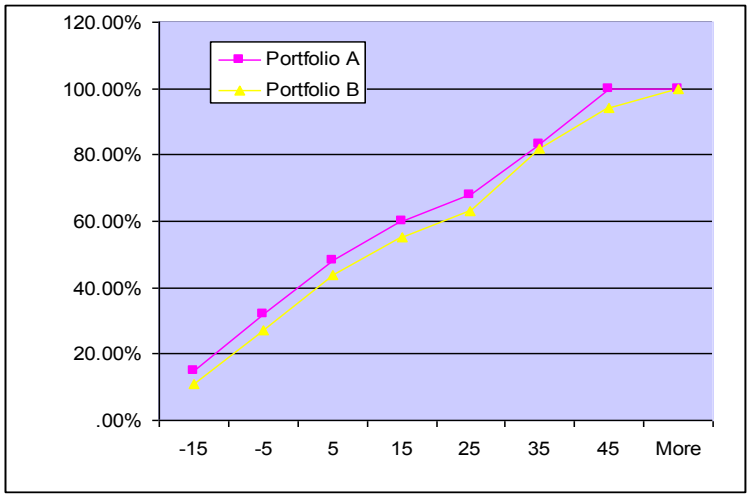
To create a Histogram in Excel have the Data Analysis add-in installed and from the Tools menu pick Data Analysis (this is where we will find 99% of the analysis tools we need).

Cumulative Data: lists a percentage for each bin which represents the percent of data which is represented up to that point (bin). The last bin will have a cumulative percentage of 100%. (See next page).

Displaying Numerical Data







The bottom histogram combines the cumulative percentages of the upper two graphs. Lets say the bins represent the quarterly return on each of the investment plans. We see that some quarters were not so good, -15% return. Other quarters were better showing up to 45% return.

The question is: which investment portfolio offers the better investment, A or B?
The key is to make money. What are the graphs telling us? We can see that one is losing less money than the other. Portfolio B is loosing less money than A.

For instance: Portfolio A has a 60% chance of returning 15% **or less** while Portfolio B has only a 57% chance of returning 15% **or less**. Most investors are going to want Portfolio B.

The most important thing is to properly interpret what the graph is telling us.

Bienayme-Chebyshev Rule

For any data set, regardless of shape, the percentage of observations contained within distances of k standard deviations around the mean must be at least

$$(1-\frac{1}{k^2})*100\%$$

Empirical Rule

In bell shaped distributions approximately 68% of the observations are contained within a distance of +/- 1 standard deviation around the mean, approximately 95% of the observations are contained within a distance of +/- 2 standard deviations around the mean, and approximately 99.7% of the observations are contained within a distance of +/- 3 standard deviation around the mean.

$$(\mu-\sigma,\mu+\sigma)=68\% \text{ of data}$$
$$(\mu-2\sigma,\mu+2\sigma)=95\% \text{ of data}$$
$$(\mu-3\sigma,\mu+3\sigma)=99.7\% \text{ of data}$$

Three Major Properties of Numerical Data:

- Central Tendency
- Variation
- Shape

Five-Number Summary

$X_{smallest}$ Q_1 Median Q_3 $X_{largest}$

If the data are perfectly symmetrical, the relationship among the various measures in the five-number summary is expressed as:

- The distance from $X_{smallest}$ to the median equals the distance from the median to $X_{largest.}$
- The distance from $X_{smallest}$ to the Q_1 equals the distance from the Q_3 to $X_{largest.}$
- The diatance from Q_1 to the median equals the distance from the median to Q_3.