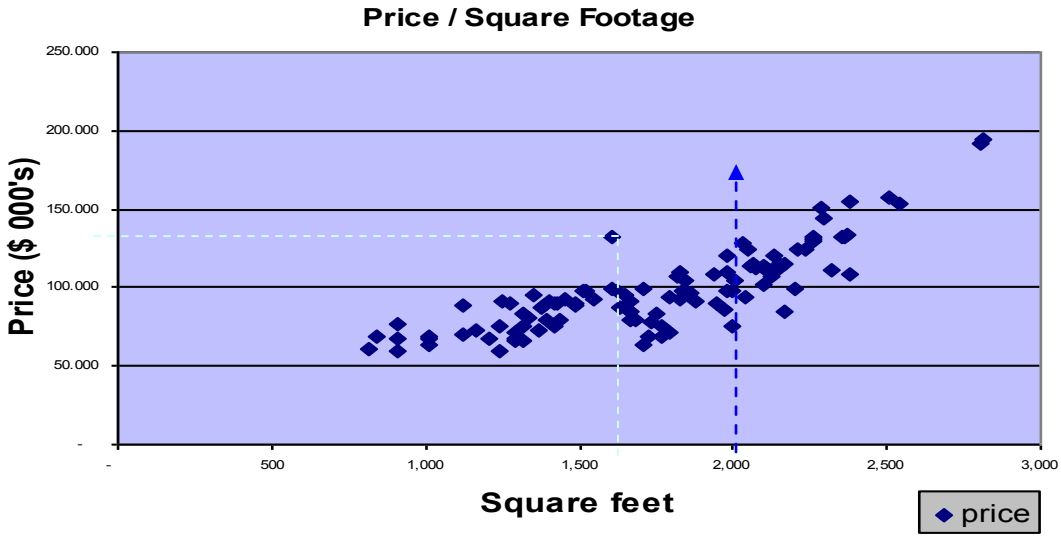
Displaying Bivariate Numerical Data



(Each point is a house corresponding to the listed square footage and price).

Scatter Diagram: A tool used to examine possible relationships between two numerical variables.

Contingency Table: a two-way table used to simultaneously study the responses of two categorical variable.

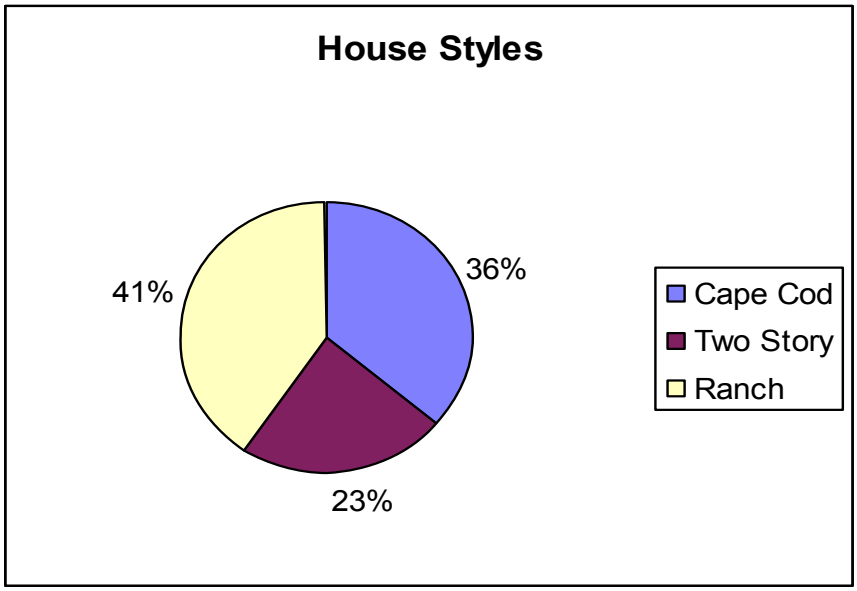
The scatter diagram and contingency table are useful tools for examining data for numerical relationships (correlation) between two variables. The above plot gives an indication of a positive correlation between the x and y variables square-footage and price. In this case each point represents a house's price vs square footage.

The closer together the points are the stronger the correlation. Widely scattered points indicate little or no correlation. Still, any given point on the x axis may have many corresponding points on the y axis (such as the number of houses at or around 2000 square feet).

A left to right decreasing relation indicates a negative relationship (slope). For example, a cars gas mileage vs weight. As the cars weight increases it's gas mileage decreases.

Displaying Categorical Data: Pie Chart

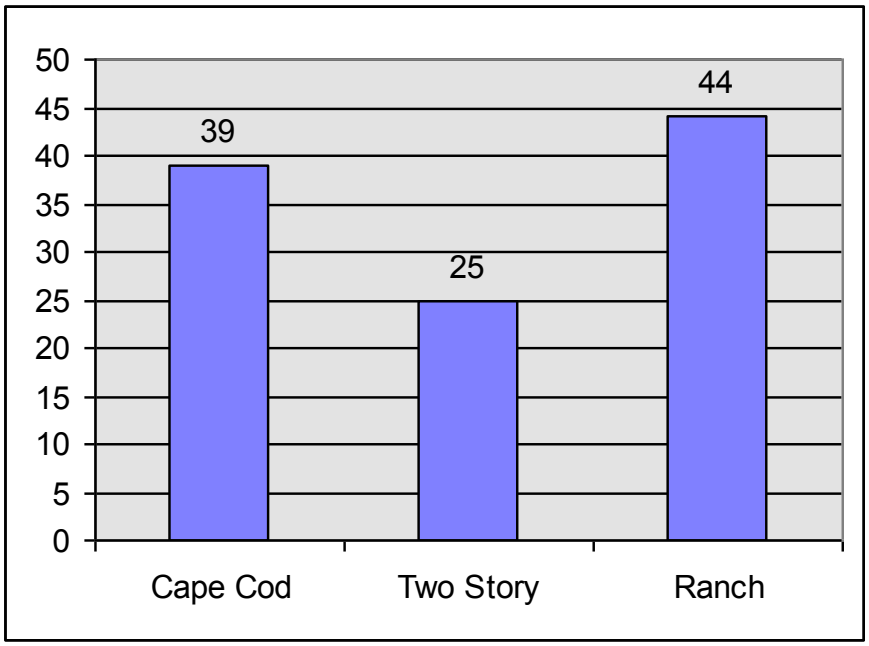
- Example: bar chart, pie chart, Pareto diagram



Categorical Data has it's own set of tools for graphical analysis. But the ways to display categorical data is limited. **Can only be presented as each categories percentage of the whole**. The percent of one value compared to another.

Displaying Categorical Data: Bar Chart

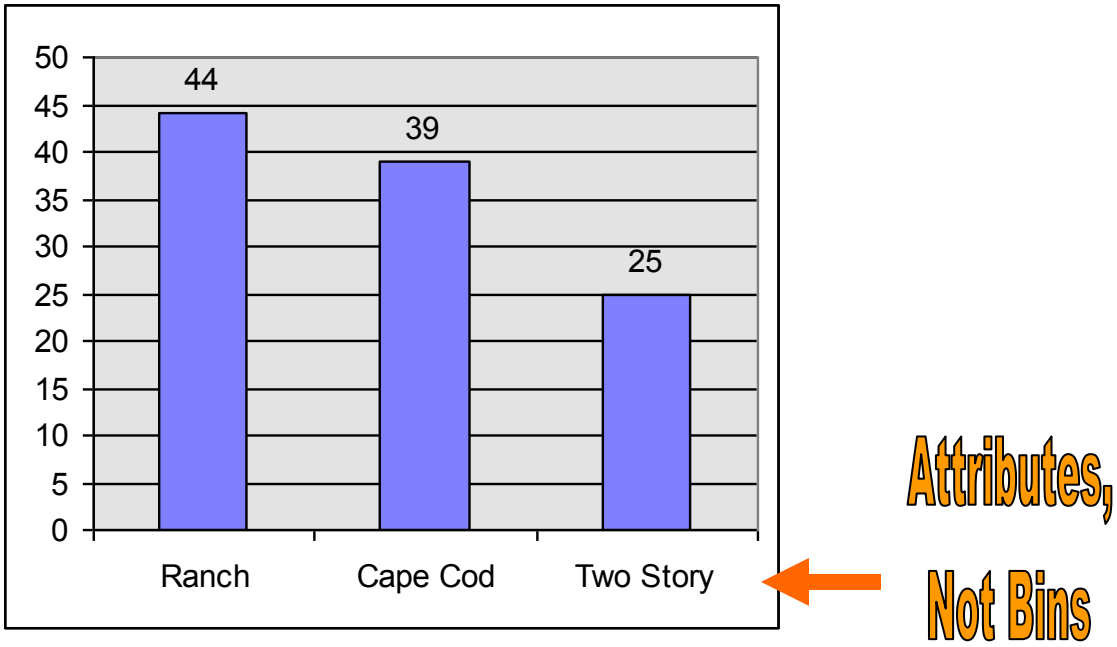
- Example: bar chart, pie chart, Pareto diagram



Summary Table: sums the data according to its category in preparation for constructing the graphical representation.

Bar chart also works with categorical data. The height of the bar represents the percentage of the population with that attribute.

Displaying Categorical Data: Pareto



Main reason to use a Pareto Diagram is to be able to graphically present the data in a way which distinguishes the "**vital few**" from the "**trivial many**." Very effective when the categorical variable of interest contains many categories.

Begin making a Pareto Diagram by constructing a Summary Table in descending order according to frequency. If necessary an "other" or "miscellaneous" category may be used but it must be placed to the extreme right of the vertical axis. **A Pareto chart is ordered with the greatest frequency category to the left**.

In Pareto format the highest bar is listed first (to the left) and follow in decending order.

Note that the x-axis for categorical data is NOT ordered and is NOT frequency as it is with numerical data.

Displaying Bivariate Categorical Data

- Contingency table

Count of style	basement		
style	0	1	Grand Total
0	14	25	39
1		25	25
2	3	41	44
Grand Total	17	91	108

Count of style	style			
basement	0	1	2	Grand Total
0	14		3	17
1	25	25	41	91
Grand Total	39	25	44	108

Count of style	basement		
style	0	1	Grand Total
0	35.90%	64.10%	100.00%
1	0.00%	100.00%	100.00%
2	6.82%	93.18%	100.00%
Grand Total	15.74%	84.26%	100.00%

In this example having a basement is the Yes/No contingency with style as the varying parameter.

Can do percentages by row, for instance Style 0, Cape, has 14/39=35.9% with basements. **108 is the total houses**. The other totals are by row and column.

When considering weather or not two variables are correlated begin by creating a contingency table. Here we examine the relationship between basement and architectural style. Style 1 has a 100% correlation for having a basement while style 2 has a 93.2% correlation.

Style	
0	Cape Cod
1	Two-Story
2	Ranch

Basement	
0	None
1	Has One

EXCEL Tutorial: PivotTables

- Here's a PDF document

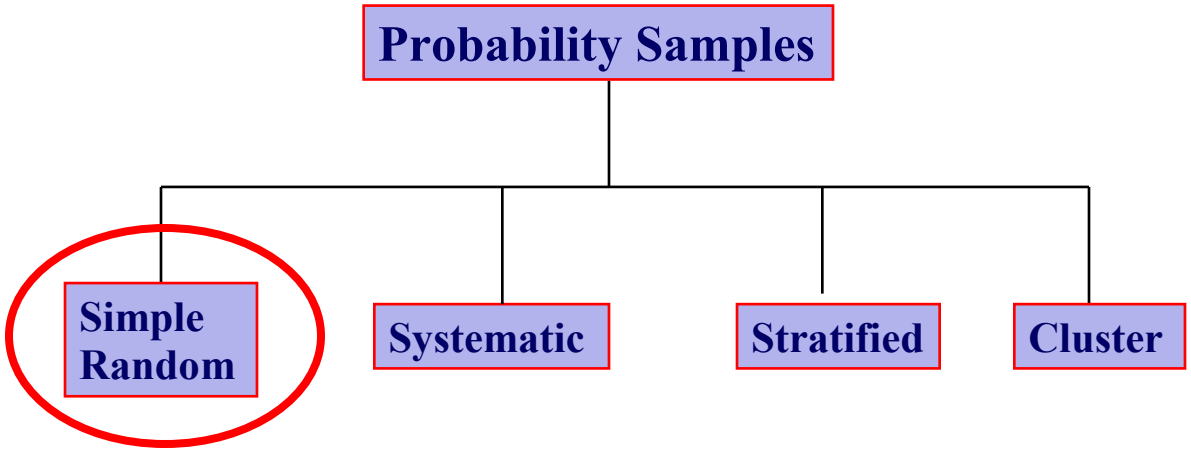
Producing Data

- Sampling methods
  THIS COURSE ALWAYS ASSUMES WE ARE USING SAMPLE DATA.

- Survey Errors

Probability Sampling

- Subjects of the sample are chosen based on known probabilities



IT IS ALWAYS ASSUMED IN THIS COURSE THAT WE ARE USING SIMPLE RANDOM

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

Difficult to obtain a true probability sample but should work toward obtaining such and acknowledging any potential biases that may exist.

Must be careful to avoid bias in sample, random sampling helps avoid this. Taking a large enough sample will help cancel out bias.

4 Most Common Types of Probability Sample:

- Simple Random
- Systematic (not really probability)
- Stratified
- Cluster Samples

A **non-probability sample** is of less use. For instance, you may just select the first 5 people in the population that you see for their opinion. But there may have been an underlying bias related to why you saw those 5 people first. So the sample does not properly represent the population and the results of the survey are skewed.

Simple Random Samples

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement or without replacement
- Samples obtained from table of random numbers or computer random number generators

n = sample size
N = frame size
Every item in the frame is numbered from 1 to N. The chance of any particular item being selected on the first draw is 1/N. Samples selected with and without replacement.

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. Probability any particulat item is selected first is 1/N. Probability any particular item is selected second is 1/(N-1), third 1/(N-2) and so on.

Fishbowl selection is not commonly used due to the difficulty in mixing the samples and of ensuring a truly random pull.

Most sampling is done WITH replacement. That means an individual could be selected two or more times. But typically the populations we are working with are so large that the probability of double selection is negligible and therefore has no influence on the results.

Random Samples

TABLE 1.1
Using a table of random numbers

		Column							
	Row	00000 12345	00001 67890	11111 12345	11112 67890	22222 12345	22223 67890	33333 12345	33334 67890
	01	49280	88924	35779	00283	81163	07275	89863	02348
	02	61870	41657	07468	08612	98083	97349	20775	45091
	03	43898	65923	25078	86129	78496	97653	91550	08078
	04	62993	93912	30454	84598	56095	20664	12872	64647
	05	33850	58555	51438	85507	71865	79488	76783	31708
Begin	06	97340	03364	88472	04334	63919	36394	11095	92470
selection	07	70543	29776	10087	10072	55980	64688	68239	20461
(row 06, column 05)	08	89382	93809	00796	95945	34101	81277	66090	88872
	09	37818	72142	67140	50785	22380	16703	53362	44940
	10	60430	22834	14130	96593	23298	56203	92671	15925
	11	82975	66158	84731	19436	55790	69229	28661	13675
	12	39087	71938	40355	54324	08401	26299	49420	59208
	13	55700	24586	93247	32596	11865	63397	44251	43189
	14	14756	23997	78643	75912	83832	32768	18928	57070
	15	32166	53251	70654	92827	63491	04233	33825	69662
	16	23236	73751	31888	81718	06546	83246	47651	04877
	17	45794	26926	15130	82455	78305	55058	52551	47182
	18	09893	20505	14225	68514	46427	56788	96297	78822
	19	54382	74598	91499	14523	68479	27686	46162	83554
	20	94750	89923	37089	20048	80336	94598	26940	36858
	21	70297	34135	53140	33340	42050	82341	44104	82949
	22	85157	47954	32979	26575	57600	40881	12250	73742
	23	11100	02340	12860	74697	96644	89439	28707	25815
	24	36871	50775	30592	57143	17381	68856	25853	35041
	25	23913	48357	63308	16090	51690	54607	72407	55538

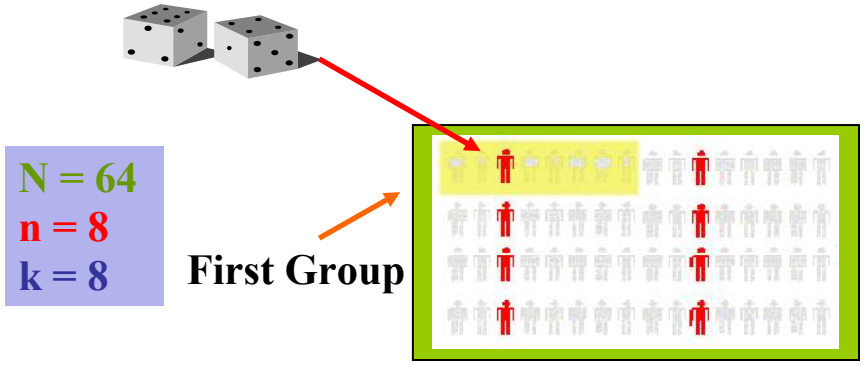
Source: Partially extracted from The Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates* (Glencoe, IL: The Free Press, 1955) and displayed in Table E.1 in Appendix E of this book.

(class example used excel "Lookup" function)

In Excel:
Tools | Data Analysis | Sampling
Input range should be a number.
Periodic is not a probability sample (systematic).
Use "Random". Taking 1 sample of size 10 (people).
Excel uses a sample-with-replacement algorithm. This could give us a duplicate. Can remedy this by taking a larger sample than you need and taking the first X unique values.

Systematic Samples

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group
- Select every k-th individual thereafter



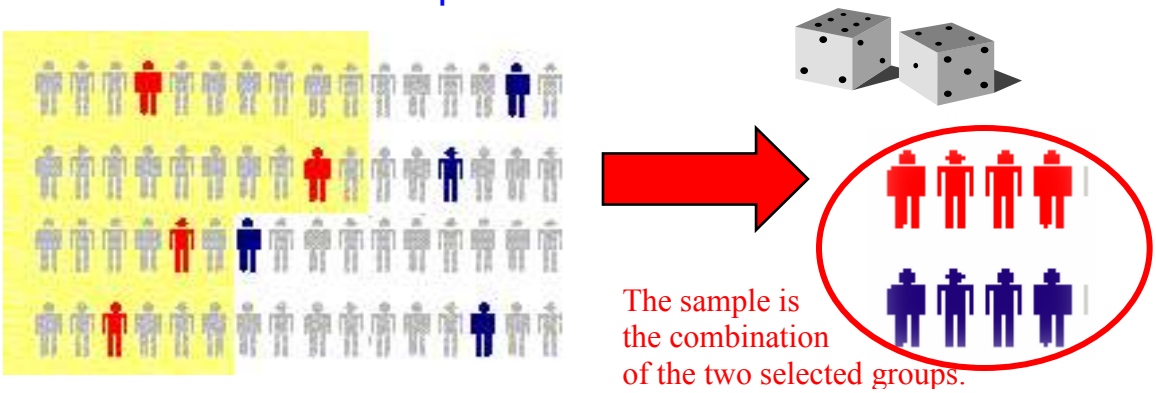
Systematic Sample: the N items in the frame are partitioned into k groups through division by the sample size n. k = N/n

Simple Random Sampling and Systematic Sampling, while simple and convenient, are less efficient than other more sophisticated methods and are susceptible to misrepresenting the populations underlying characteristics (parameters).

The systematic method is not preferred.

Stratified Samples

- Population divided into two or more groups (called **strata**) according to some common characteristic.
- Simple random sample selected from each group
- The two or more samples are combined into one



This method ensures representation of items from across the entire population. This ensures a greater precision in the estimates of the underlying population parameters.

This method can be used to help insure all members of a population are represented in a sample. For instance, if there is a small percentage of women in the population use this method to create strata and than sample each strata. In the above the yellow background people are men and the white background people are women.

Another example would be to break up income classes into strata.

This method is more sophisticated than systematic method but also more complicated to analysis.

Advantages and Disadvantages

- Simple random sample and systematic sample
  - Simple to use
  - May not be a good representation of the population's underlying characteristics
- Stratified sample
  - Ensures representation of individuals across the entire population
- Cluster sample
  - More cost effective
  - Less efficient (need larger sample to acquire the same level of precision)

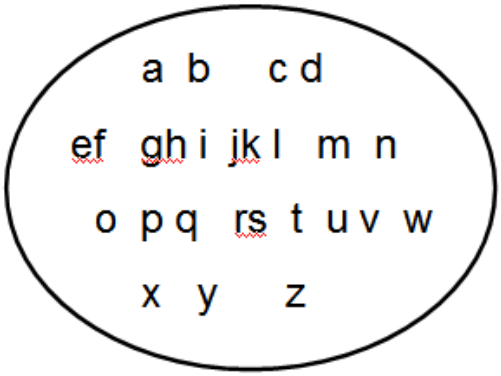
Cluster Sample: N items in a frame are divided into several clusters so that each cluster is representative of the entire population. Natural examples include countries, election districts, city blocks, apartment buildings, and families. Must find a group with the same characteristics as the whole population and sample that group. Can be a dangerous method. For instance, consider the possibility of finding a town with a population which ahs the same characteristics as the whole country. Drawback is that some of the population may not be represented.

Key Definitions

- A population (universe) is the collection of things under consideration
- A sample is a portion of the population selected for analysis (sample is not equal to the whole population)
- A **parameter** is a summary measure computed to describe a characteristic of the **population**
- A **statistic** is a summary measure computed to describe a characteristic of the **sample**

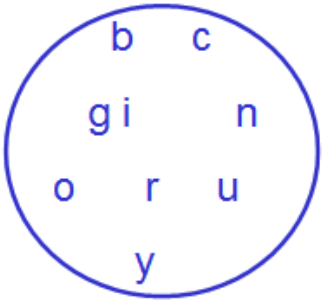
A parameter will give an exact result because it measures the entire population.
A statistic will not be exact but we can measure how far off the we expect the statistic to be.

Population



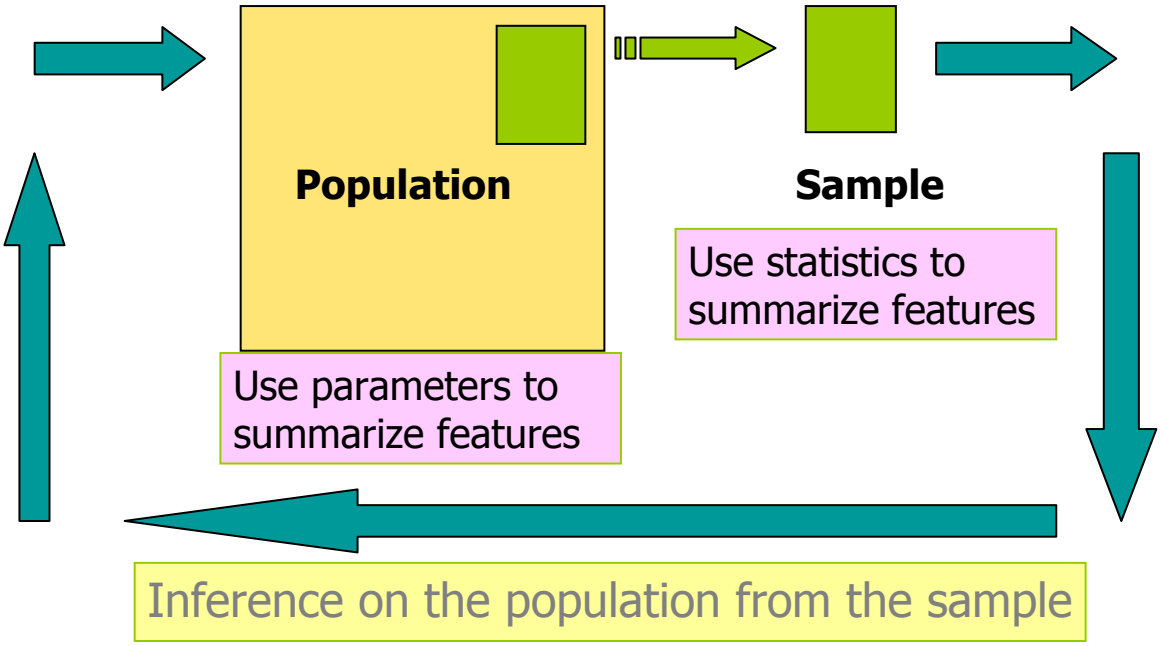
Measures used to describe the population are called parameters

Sample



Measures computed from sample data are called statistics

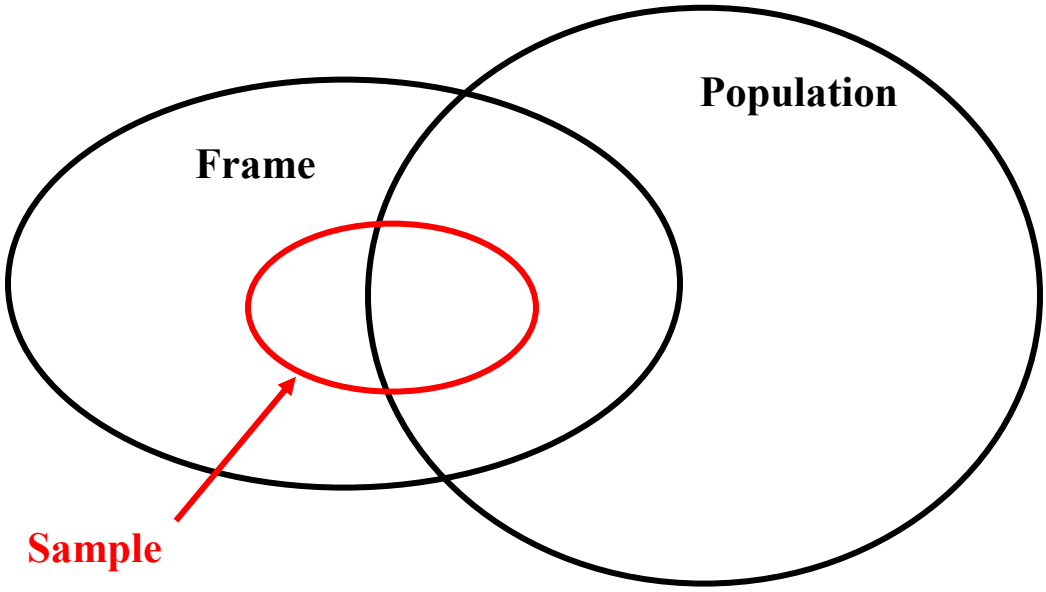
Population and Sample



Statistics will tell us exactly how close we are coming to the population parameters.

The keys here are Population and Sample. For example, a Parameter would be the average salary (income) of a US household. To find this **PARAMETER** we would have to know the income of the entire population. This is impractical so we take a sample of several thousand and we compute the average of the sample data. This result is called a **STATISTIC**. Statistics are computed from sample data, parameters from populations. We are interested in the Parameter but it is not always possible to derive the parameter. So we use statistics and if done properly the statistic will give us results very close to the parameter.

Venn Diagram



We need a good way to select from the population. Can end up with parts of the population which do not make it into the frame and visa-versa. For example, you will not want to poll a 3 year old but he or she may make it to the frame. Ideally the frame and population would be exactly the same, but this will not be possible. The undesirable portions are the members which are only in the frame or only in the population. For example, in a phone survey the people with unlisted numbers will be part of the population but not make it into the frame.

THE SAMPLE IS ALWAYS DRAWN FROM THE FRAME.

In the ideal case the frame and population would be exactly the same.

Always ask yourself, "How was the frame selected?"

A **SAMPLE** is a portion of the whole population selected for analysis.

Reasons for Drawing a Sample

- Less time consuming than a census
- Less costly to administer than a census
- Less cumbersome and more practical to administer than a census of the targeted population

Evaluating Survey Worthiness

- What is the purpose of the survey?
- Is the survey based on a probability sample?
- Coverage error – appropriate frame
- Nonresponse error – follow up
- Measurement error – good questions elicit good responses
- Sampling error – always exists

Can we still get a biased result even if we start with a perfect frame? Yes, for these reasons.

Types of Survey Errors

- **Coverage error**
  Exclusion from frame will bias the result.
- **Non response error**
  Creates bias, Shere Hite & Ann Landers examples.
- **Sampling error**
  Always exist because cannot access entire population.
- **Measurement error**

Excluded from frame.

Follow up on non responses.

Chance differences from sample to sample.

Bad Question!

Statistics will identify the sampling error. Sampling error will always exist because we are not going to interview or measure the entire population. The results (statistics) will vary from sample to sample. But statistics is able to quantify this error. Nullifying this sampling error makes up the bulk of the course.

Voluntary Response (anyone who wants to respond can respond) surveys are completely unscientific. No value whatsoever.

Measurement Errors

- Question Phrasing, Avoid "Negation Phrasing"
- Telescoping effect:. People will tend to believe they experienced a recent event more recently than in actual fact. Like, "when was your last dentist visit?"
- Halo Effect: subject will lie if he or she prefers to be viewed as a "good, socially acceptable" person.
- Overzealous / Under zealous: subject wants to influence study.

Think about these things when designing a questionnaire.
Notes from Reading:

Sections 2.1 – 2.5

Ordered Array: an ordered sequence from smallest to largest. Facilitates identifying lowest and highest values, concentrations of data, typical values, and where the majority of values are concentrated. The data can be organized into a stem-and-leaf display in order to study it's characteristics.

A **stem-and-leaf** display separates data entries into leading digits, or stems, and trailing digits, of leafs. Example, in the number 10.9 the value 10 is the stem and 9 is the leaf.

Ex. Money spent at fast food restaurant:

Raw Data

5.35 4.75 4.3 5.47 4.85 6.62 3.54 4.87 6.26 5.48 7.27 8.45 6.05 4.76 5.91

Ordered Array

3.54 4.3 4.8 4.76 4.85 4.87 5.35 5.47 5.48 5.91 6.05 6.26 6.62 7.27 8.45

Fast Food Stem & Leaf **Stem unit: 1**

Statistics	
Sample Size	15.00
Mean	5.60
Median	5.47
Std. Deviation	1.23
Minimum	3.54
Maximum	8.45

3	5
4	3 8 8 9 9
5	4 5 5 9
6	0 3 6
7	3
8	4

Graph shows that the majority of data is clustered around 4.3 to 6.6.

Frequency Distribution: a summary table in which the data are arranged into numerically ordered class groupings or categories. Attention must be given to slecting the appropriate number of class groupings and obtaining a suitable class interval (width). In general try to limit class groupings to between 5 and 15. More or less hampers what is learned from the data.

width of class interval = range / number of class groupings

Range: the numerical difference between the largest value and the smallest value in a data set.

$$\text{Relative Frequency Distribution} = \frac{\text{Frequencies in Each Class}}{\text{Total Number of Observations}}$$

$$\text{Percentage Distribution} = \text{Relative Frequency} * 100$$

5-Year Annualized Percentage Return			Number of Funds	Proportion of Funds	Percentage of Funds
-10	to	-5	1	0.006	0.6
-5	to	0	3	0.019	1.9
0	to	5	14	0.089	8.9
5	to	10	58	0.367	36.7
10	to	15	61	0.386	38.6
15	to	20	17	0.108	10.8
20	to	25	3	0.019	1.9
25	to	30	1	0.006	0.6
	Total		158	1.000	100.0

Population Mean

The sum of the values in the population divided by the population size, N.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Population Variance

The sum of the squared differences around the population mean divided by the population size, N.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Population Standard Deviation

The population standard deviation is the square of the population variance.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$