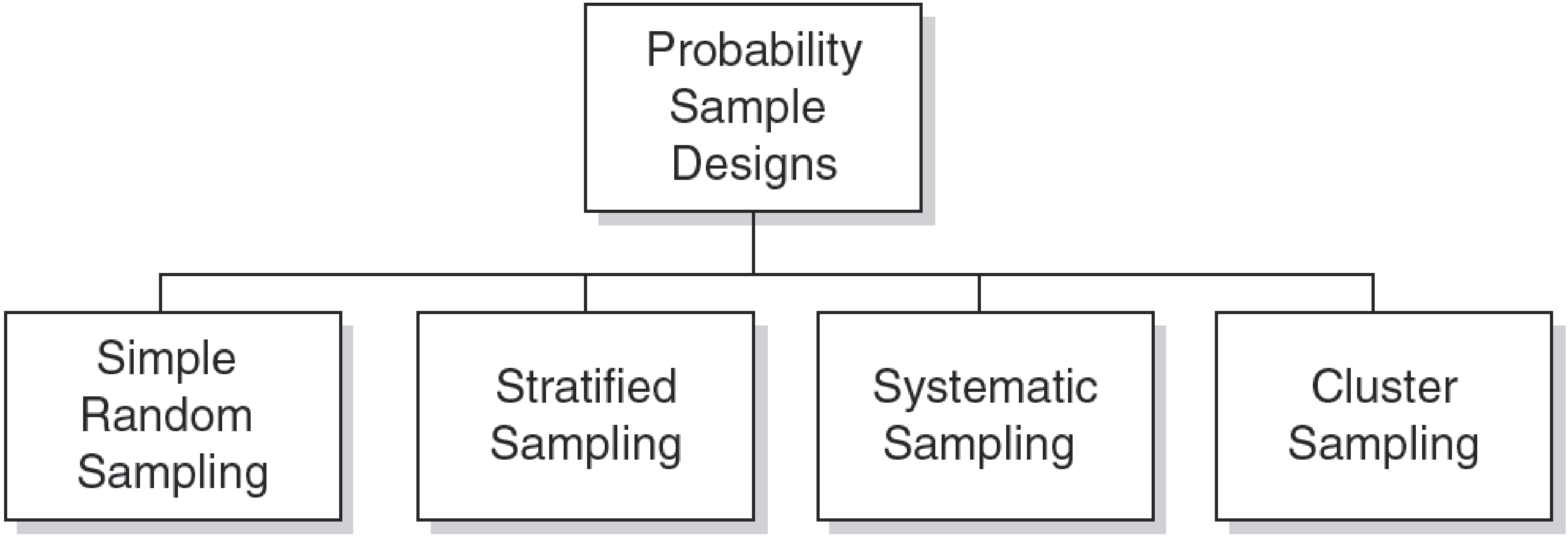
Major Types of Probability Sampling



Simple Random Sampling

Simple random sampling is a probability sampling procedure that gives every element in the target population, and each possible sample of a given size, an equal chance of being selected. As such, it is an equal probability selection method (EPSEM).

Simple Random Sampling

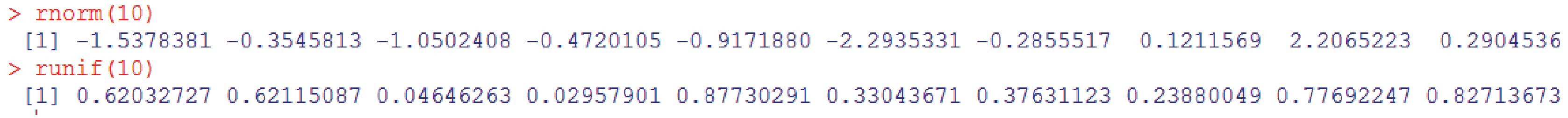
There are **six major steps** in selecting a simple random sample:

- **Step 1** Define the target population.
- **Step 2** Identify an existing sampling frame of the target population or develop a new one.
- **Step 3** Evaluate the sampling frame for undercoverage, overcoverage, multiple coverage, and clustering, and make adjustments where necessary.
- **Step 4** Assign a unique number to each element in the frame.
- **Step 5** Determine the sample size.
- **Step 6** Randomly select the predetermined number of population elements.

Random Selection Techniques

There are several strategies to "randomly select":

- Lottery method (also known as *blind draw method* or *hat method*)
- Table of random numbers
- Randomly generated numbers using a computer programme (e.g., R - see exercise classes)



```
> rnorm(10)
 [1] -1.5378381 -0.3545813 -1.0502408 -0.4720105 -0.9171880 -2.2935331 -0.2855517  0.1211569  2.2065223  0.2904536
> runif(10)
 [1] 0.62032727 0.62115087 0.04646263 0.02957901 0.87730291 0.33043671 0.37631123 0.23880049 0.77692247 0.82713673
```

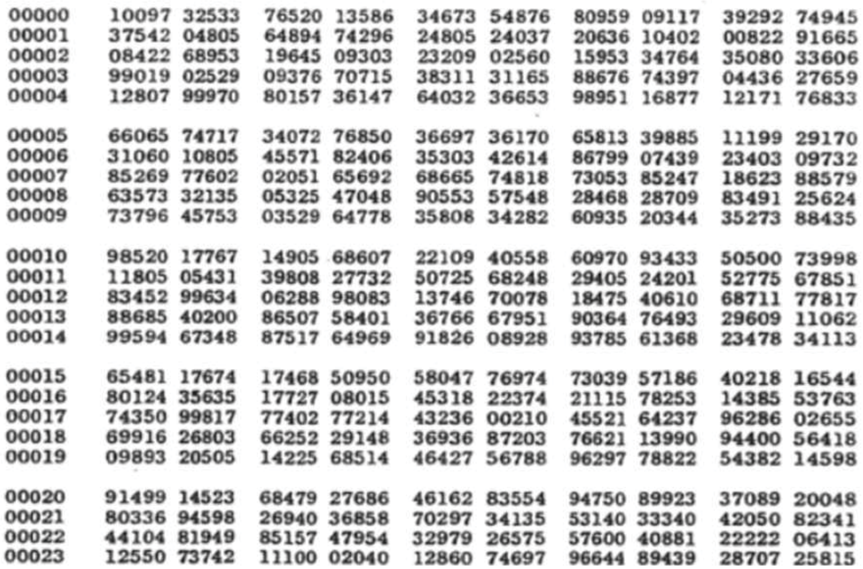
Using a Table of Random Numbers

How to randomly choose n individuals from a group of N?

- We first label each of the N individuals with a number (typically from 1 to N, or 0 to $N-1$)
- A list of random digits is parsed into digits of the same length as N (if $N=233$, then its length is 3; if $N=18$, its length is 2).
- The parsed list is read in sequence and the first n entries from this list, corresponding to a label in our group of N, are selected.
- The n individuals with these labels constitute our selection.

Using a Table of Random Numbers

Part of a random number table:

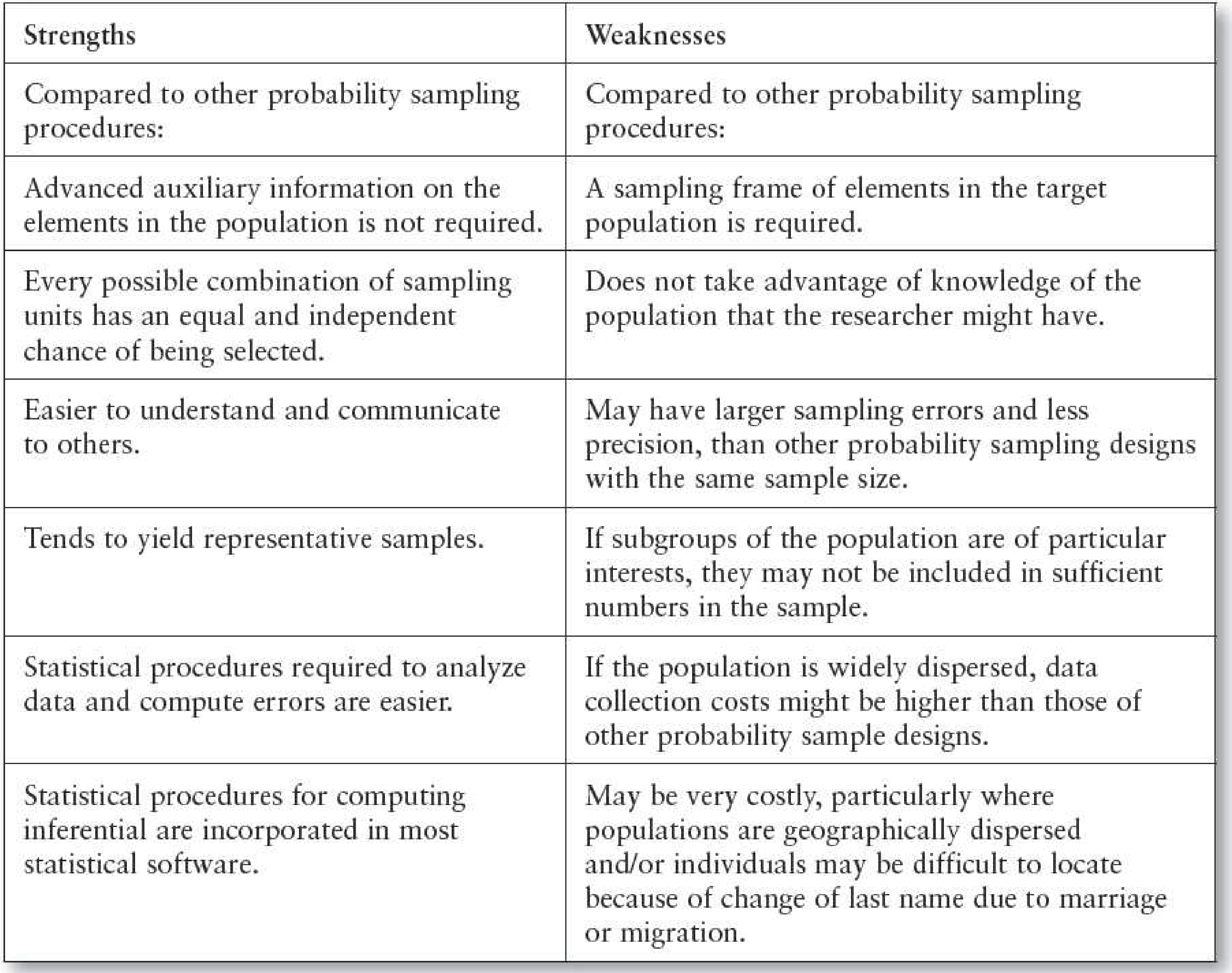


00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815

Subtypes of Simple Random Sampling

- **Sampling with Replacement** In sampling with replacement, after an element has been selected from the sampling frame, it is returned to the frame and is eligible to be selected again
- **Sampling without Replacement** In sampling without replacement, after an element is selected from the sampling frame, it is removed from the population and is not returned to the sampling frame. Sampling without replacement tends to be more efficient than sampling with replacement in producing representative samples. It does not allow the same population element to enter the sample more than once. [*Here: When talking about Simple Random Sampling, we will mean sampling without replacement, unless stated otherwise.*]

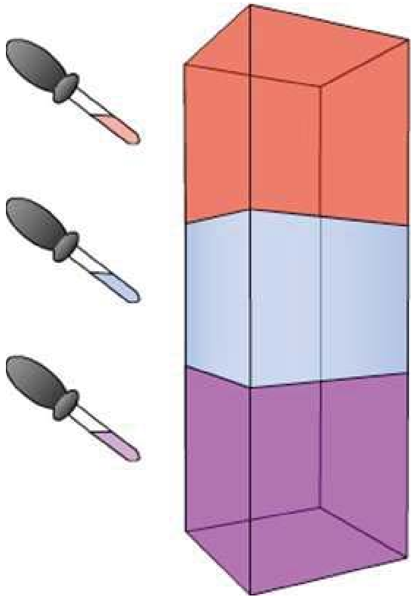
Stength and Weaknesses of Simple Random Sampling



Strengths	Weaknesses
Compared to other probability sampling procedures:	Compared to other probability sampling procedures:
Advanced auxiliary information on the elements in the population is not required.	A sampling frame of elements in the target population is required.
Every possible combination of sampling units has an equal and independent chance of being selected.	Does not take advantage of knowledge of the population that the researcher might have.
Easier to understand and communicate to others.	May have larger sampling errors and less precision, than other probability sampling designs with the same sample size.
Tends to yield representative samples.	If subgroups of the population are of particular interests, they may not be included in sufficient numbers in the sample.
Statistical procedures required to analyze data and compute errors are easier.	If the population is widely dispersed, data collection costs might be higher than those of other probability sample designs.
Statistical procedures for computing inferential are incorporated in most statistical software.	May be very costly, particularly where populations are geographically dispersed and/or individuals may be difficult to locate because of change of last name due to marriage or migration.

Stratified Sampling

Stratified Sampling is a probability sampling procedure in which the target population is first separated into mutually exclusive, homogeneous segments (strata), and then a simple random sample is selected from each segment (stratum). The samples selected from the various strata are then combined into a single sample.



Stratified Sampling

There are **eight major steps** in selecting a simple random sample:

- **Step 1** Define the target population.
- **Step 2** Identify stratification variable(s) and determine the number of strata to be used. The stratification variables should relate to the purposes of the study. If the purpose of the study is to make subgroup estimates, the stratification variables should be related to those subgroups. The availability of auxiliary information often determines the stratification variables that are used. Considering that as the number of stratification variables increases, the likelihood increases that some of the variables will cancel the effects of other variables, not more than four to six stratification variables and not more than six strata for a particular variable should be used.

Stratified Sampling

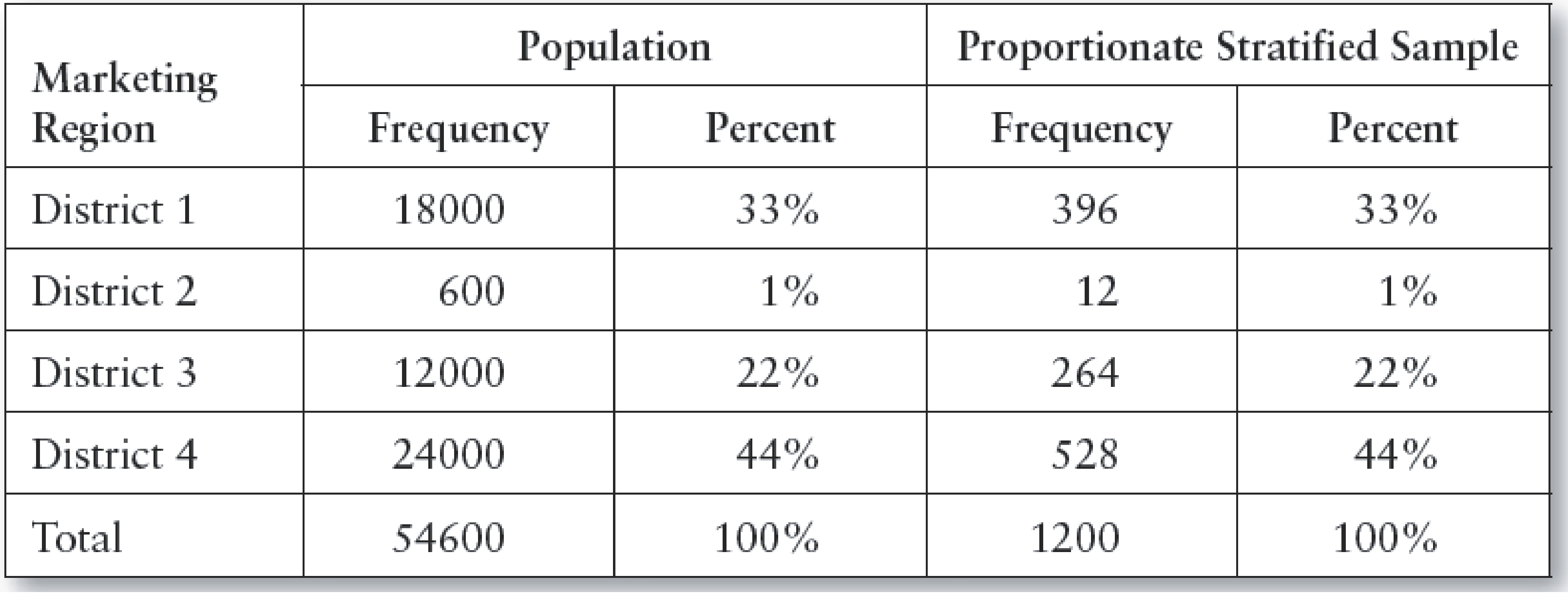
- **Step 3** Identify an existing sampling frame or develop a sampling frame that includes information on the stratification variable(s) for each element in the target population. If the sampling frame does not include information on the stratification variables, stratification would not be possible.
- **Step 4** Evaluate the sampling frame for undercoverage, overcoverage, multiple coverage, and clustering, and make adjustments where necessary.
- **Step 5** Divide the sampling frame into strata, categories of the stratification variable(s), creating a sampling frame for each stratum. Within-stratum differences should be minimized, and between-strata differences should be maximized. The strata should not be overlapping, and altogether, should constitute the entire population. The strata should be independent and mutually exclusive subsets of the population. Every element of the population must be in one and only one stratum.

Stratified Sampling

- **Step 6** Assign a unique number to each element.
- **Step 7** Determine the sample size for each stratum. The numerical distribution of the sampled elements across the various strata determines the type of stratified sampling that is implemented. It may be a proportionate stratified sampling or one of the various types of disproportionate stratified sampling.
- **Step 8** Randomly select the targeted number of elements from each stratum. At least one element must be selected from each stratum for representation in the sample; and at least two elements must be chosen from each stratum for the calculation of the margin of error of estimates computed from the data collected.

Subtypes of Stratified Sampling

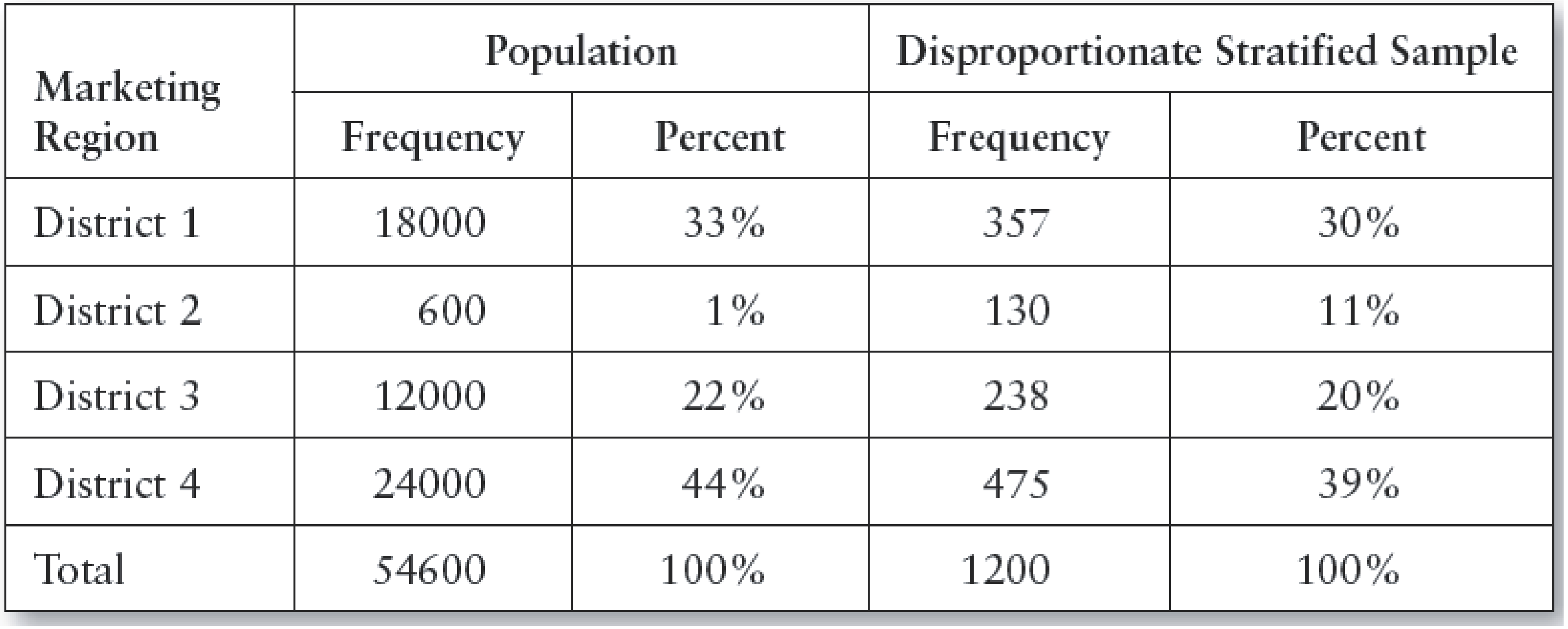
- **Proportionate Stratified Sampling** In proportionate stratified sampling, the number of elements allocated to the various strata is proportional to the representation of the strata in the target population. This sampling procedure is used when the purpose of the research is to estimate a population parameter.



Marketing Region	Population		Proportionate Stratified Sample	
	Frequency	Percent	Frequency	Percent
District 1	18000	33%	396	33%
District 2	600	1%	12	1%
District 3	12000	22%	264	22%
District 4	24000	44%	528	44%
Total	54600	100%	1200	100%

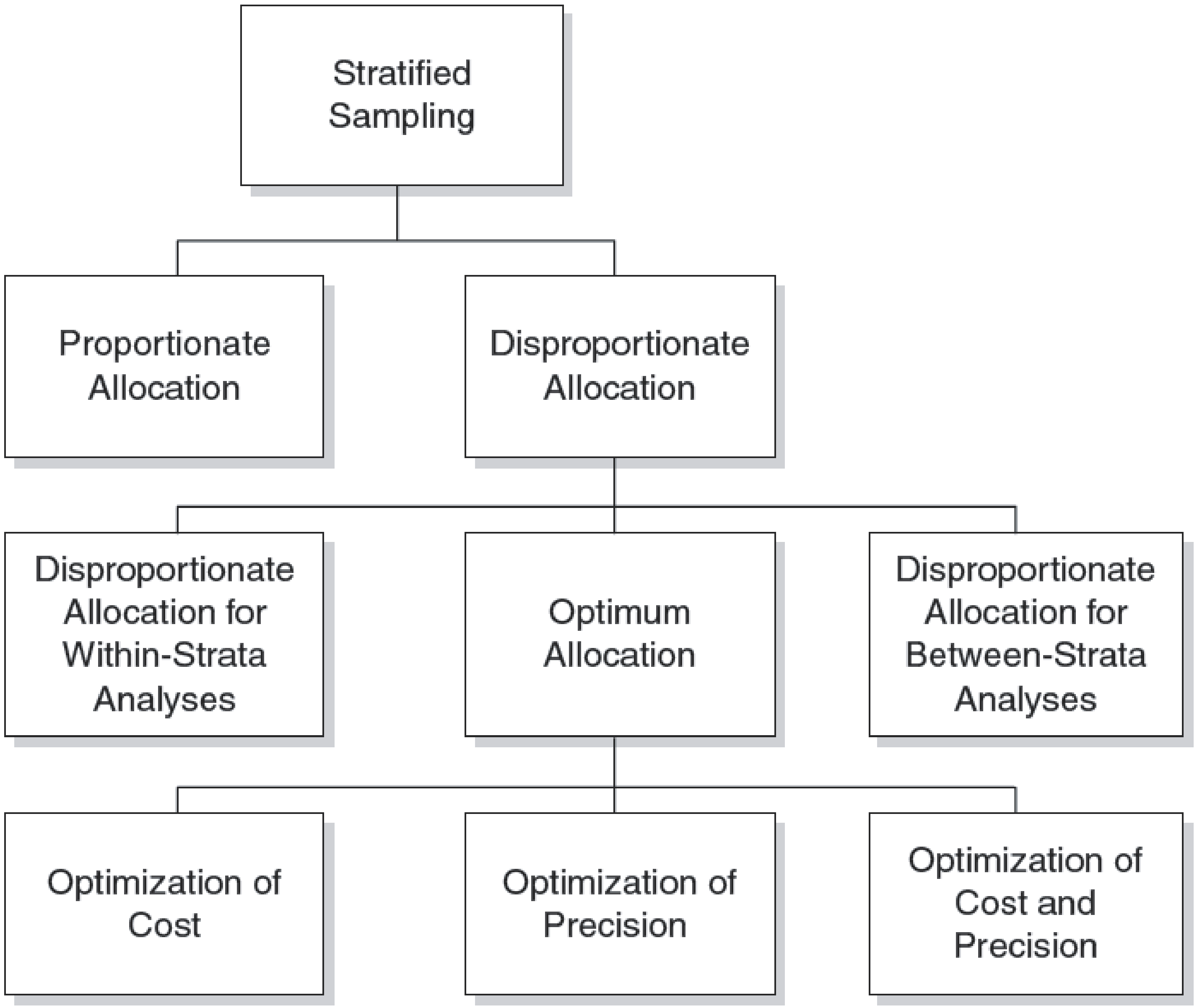
Subtypes of Stratified Sampling

- **Disproportionate Stratified Sampling** Disproportionate stratified sampling is a stratified sampling procedure in which the number of elements sampled from each stratum is not proportional to their representation in the total population. Population elements are not given an equal chance to be included in the sample. The same sampling fraction is not applied to each stratum.

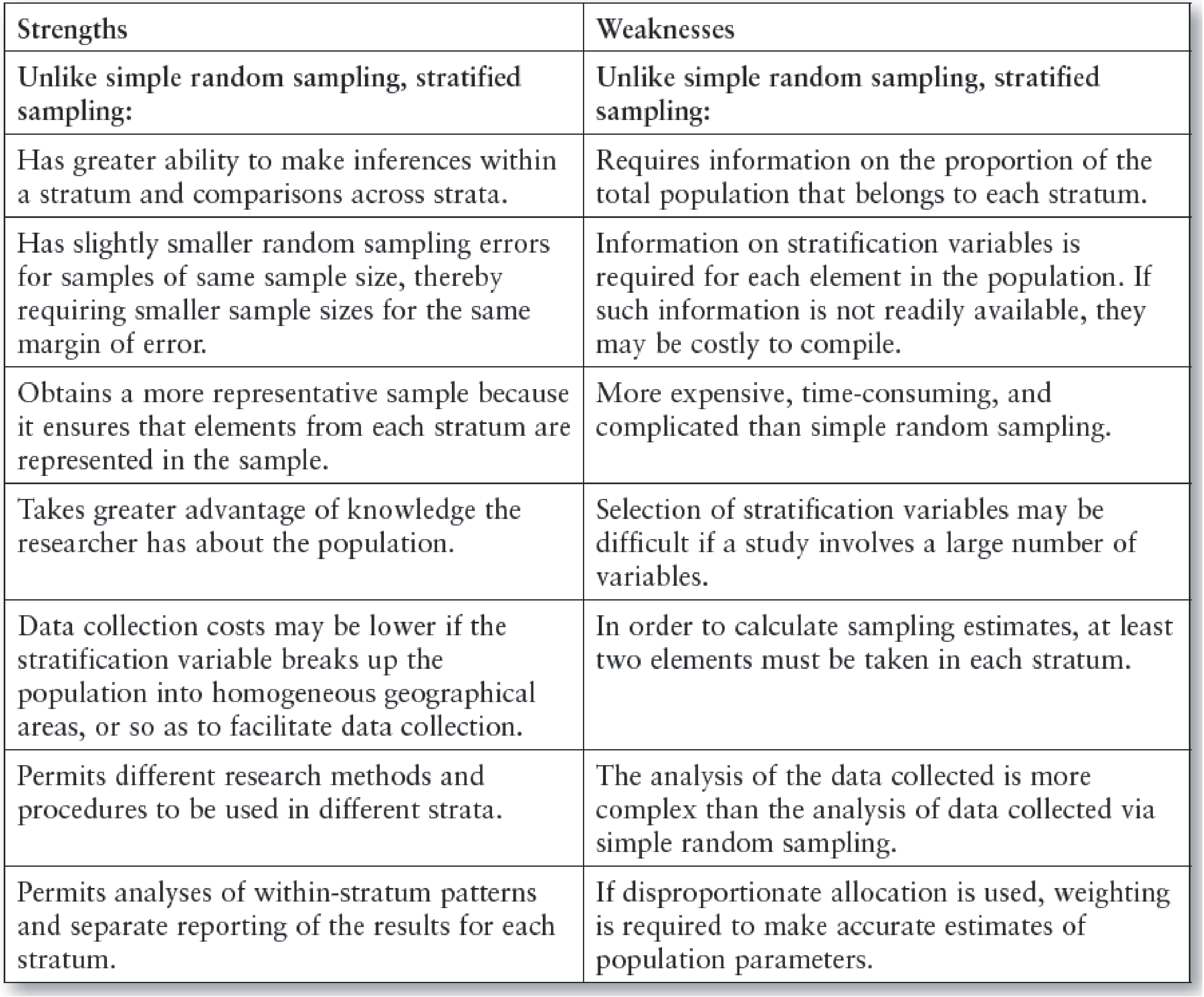


Marketing Region	Population		Disproportionate Stratified Sample	
	Frequency	Percent	Frequency	Percent
District 1	18000	33%	357	30%
District 2	600	1%	130	11%
District 3	12000	22%	238	20%
District 4	24000	44%	475	39%
Total	54600	100%	1200	100%

Subtypes of Stratified Sampling

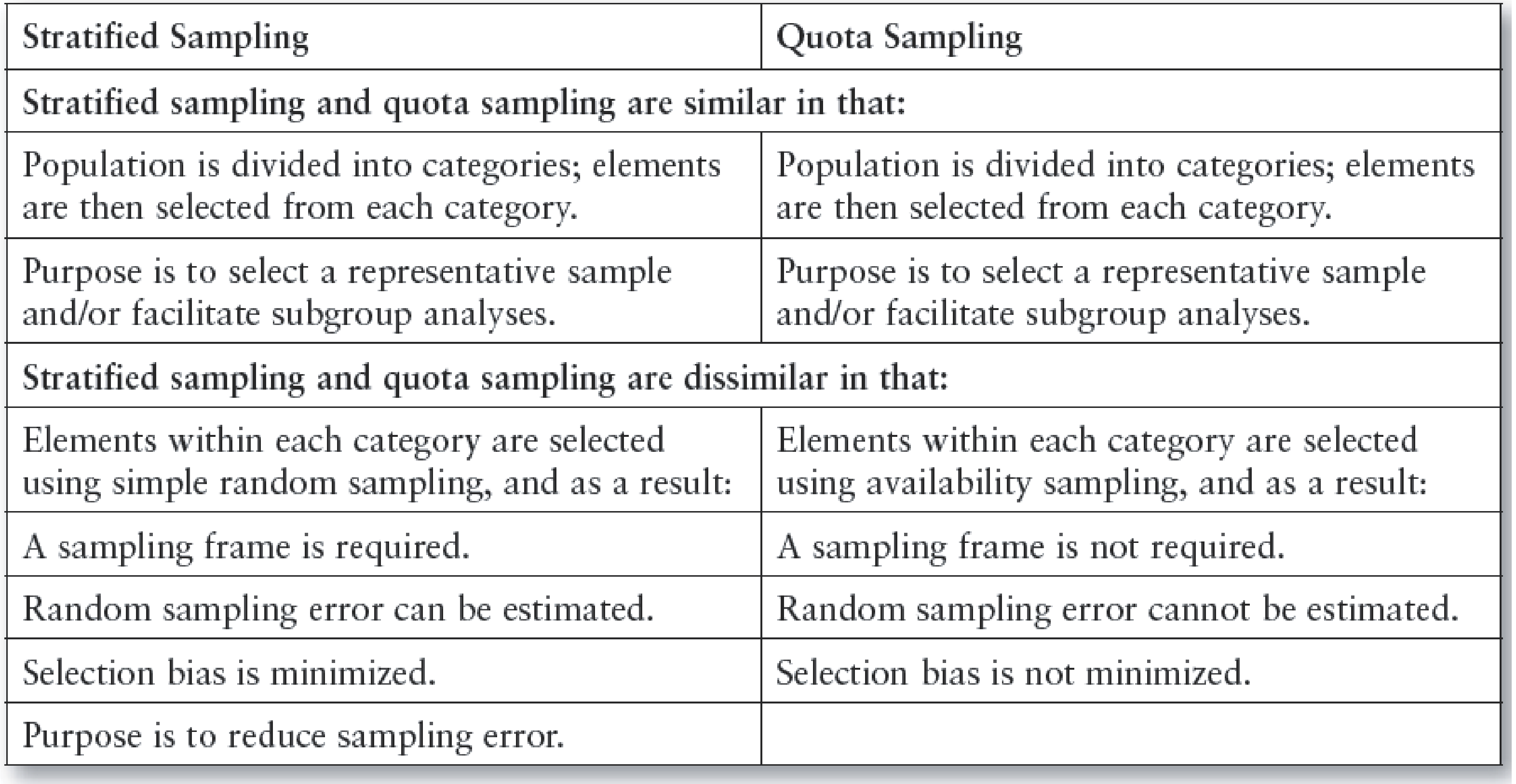

Stratified Sampling
Proportionate Allocation
Disproportionate Allocation
Disproportionate Allocation for Within-Strata Analyses
Optimum Allocation
Disproportionate Allocation for Between-Strata Analyses
Optimization of Cost
Optimization of Precision
Optimization of Cost and Precision

Stength and Weaknesses of Stratified Sampling



Strengths	Weaknesses
Unlike simple random sampling, stratified sampling:	Unlike simple random sampling, stratified sampling:
Has greater ability to make inferences within a stratum and comparisons across strata.	Requires information on the proportion of the total population that belongs to each stratum.
Has slightly smaller random sampling errors for samples of same sample size, thereby requiring smaller sample sizes for the same margin of error.	Information on stratification variables is required for each element in the population. If such information is not readily available, they may be costly to compile.
Obtains a more representative sample because it ensures that elements from each stratum are represented in the sample.	More expensive, time-consuming, and complicated than simple random sampling.
Takes greater advantage of knowledge the researcher has about the population.	Selection of stratification variables may be difficult if a study involves a large number of variables.
Data collection costs may be lower if the stratification variable breaks up the population into homogeneous geographical areas, or so as to facilitate data collection.	In order to calculate sampling estimates, at least two elements must be taken in each stratum.
Permits different research methods and procedures to be used in different strata.	The analysis of the data collected is more complex than the analysis of data collected via simple random sampling.
Permits analyses of within-stratum patterns and separate reporting of the results for each stratum.	If disproportionate allocation is used, weighting is required to make accurate estimates of population parameters.

Differences between Quota and Stratified Sampling



Stratified Sampling	Quota Sampling
Stratified sampling and quota sampling are similar in that:	
Population is divided into categories; elements are then selected from each category.	Population is divided into categories; elements are then selected from each category.
Purpose is to select a representative sample and/or facilitate subgroup analyses.	Purpose is to select a representative sample and/or facilitate subgroup analyses.
Stratified sampling and quota sampling are dissimilar in that:	
Elements within each category are selected using simple random sampling, and as a result:	Elements within each category are selected using availability sampling, and as a result:
A sampling frame is required.	A sampling frame is not required.
Random sampling error can be estimated.	Random sampling error cannot be estimated.
Selection bias is minimized.	Selection bias is not minimized.
Purpose is to reduce sampling error.	

Systematic Sampling

Systematic Sampling Systematic sampling (or interval random sampling) is a probability sampling procedure in which a random selection is made of the first element for the sample, and then subsequent elements are selected using a fixed or systematic interval until the desired sample size is reached. The random start distinguishes this sampling procedure from its non-probability counterpart.

- For example, after a random start, one may systematically select every *i*-th patient visiting an emergency room in a hospital, store customers standing in line, or records in file drawers.
- At a technical level, systematic sampling does not create a truly random sample. It is often referred to as “pseudo random sampling”, “pseudo simple random sampling”, or “quasi-random sampling”.

Systematic Sampling

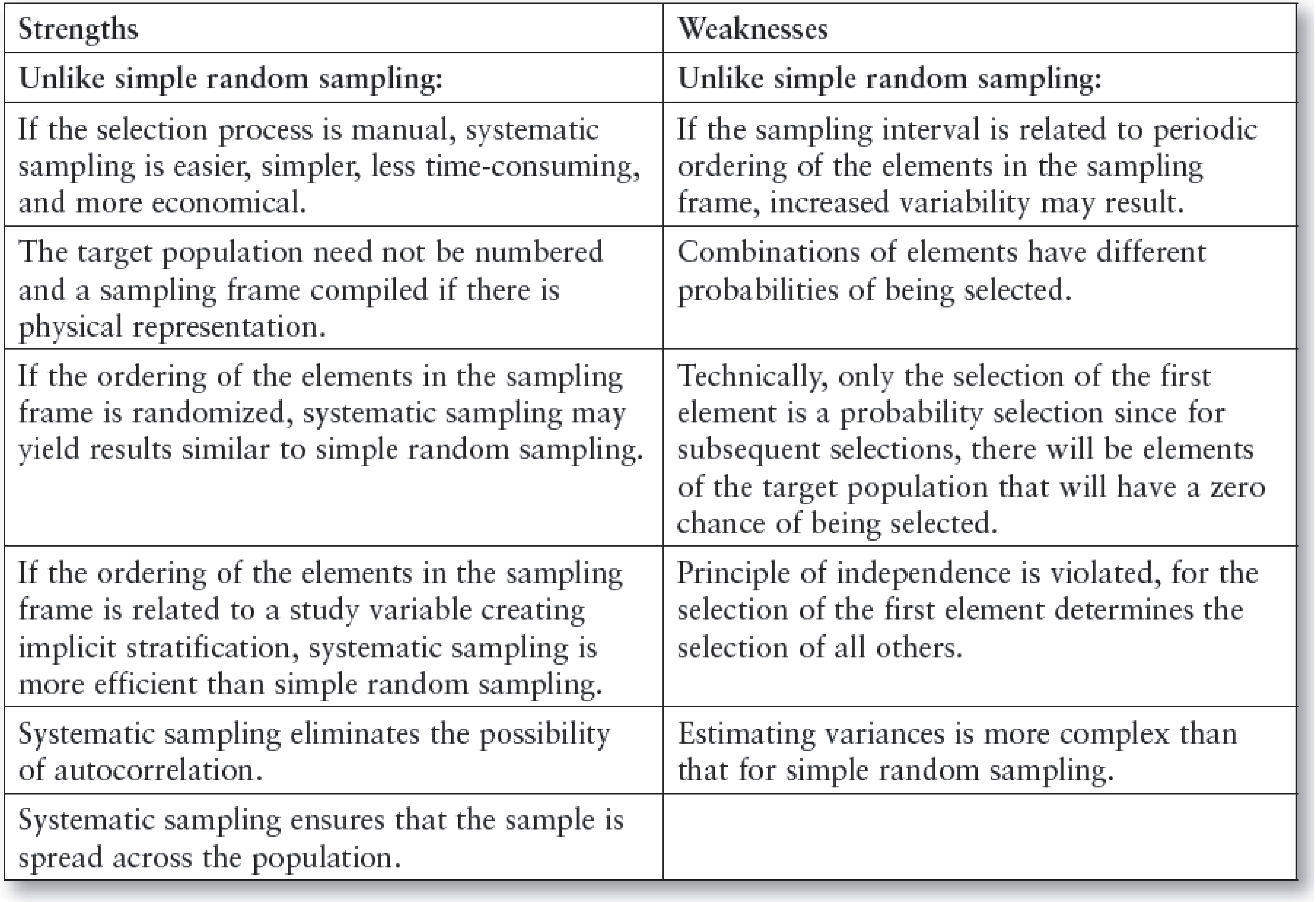
There are **eight major steps** in selecting a simple random sample:

- **Step 1** Define the target population.
- **Step 2** Determine the desired sample size (n).
- **Step 3** Identify an existing sampling frame or develop a sampling frame of the target population.
- **Step 4** Evaluate the sampling frame for undercoverage, overcoverage, multiple coverage, clustering, and periodicity, and make adjustments where necessary. Ideally, the list will be in a random order with respect to the study variable. If the sampling frame is randomized, systematic sampling is considered to be a good approximation of simple random sampling.

Systematic Sampling

- **Step 5** Determine the number of elements in the sampling frame (N).
- **Step 6** Calculate the sampling interval (i) by dividing the number of elements in the sampling frame (N) by the targeted sample size (n). One should ignore a remainder and round down or truncate to the nearest whole number. Rounding down and truncating may cause the sample size to be larger than desired. If so, one may randomly delete the extra selections. If the exact size of the population is not known and impractical to determine, one may fix the sampling fraction.
- **Step 7** Randomly select a number, r, from "1" through i.
- **Step 8** Select for the sample, $r, r + i, r + 2i, r + 3i$, and so forth, until the frame is exhausted.

Strengths and Weaknesses of Systematic Sampling



Strengths	Weaknesses
Unlike simple random sampling:	Unlike simple random sampling:
If the selection process is manual, systematic sampling is easier, simpler, less time-consuming, and more economical.	If the sampling interval is related to periodic ordering of the elements in the sampling frame, increased variability may result.
The target population need not be numbered and a sampling frame compiled if there is physical representation.	Combinations of elements have different probabilities of being selected.
If the ordering of the elements in the sampling frame is randomized, systematic sampling may yield results similar to simple random sampling.	Technically, only the selection of the first element is a probability selection since for subsequent selections, there will be elements of the target population that will have a zero chance of being selected.
If the ordering of the elements in the sampling frame is related to a study variable creating implicit stratification, systematic sampling is more efficient than simple random sampling.	Principle of independence is violated, for the selection of the first element determines the selection of all others.
Systematic sampling eliminates the possibility of autocorrelation.	Estimating variances is more complex than that for simple random sampling.
Systematic sampling ensures that the sample is spread across the population.	

Cluster Sampling

Cluster Sampling Cluster sampling is a probability sampling procedure in which elements of the population are randomly selected in naturally occurring groupings (clusters). In the context of cluster sampling, a cluster is an aggregate or intact grouping of population elements. Element sampling is the selection of population elements individually, one at a time. On the other hand, cluster sampling involves the selection of population elements not individually, but in aggregates.

Cluster Sampling

There are **six major steps** in selecting a cluster sample:

- **Step 1** Define the target population.
- **Step 2** Determine the desired sample size.
- **Step 3** Identify an existing sampling frame or develop a new sampling frame of clusters of the target population.
- **Step 4** Evaluate the sampling frame for undercoverage, overcoverage, multiple coverage, and clustering, and make adjustments where necessary. Ideally, the clusters would be as heterogeneous as the population, mutually exclusive, and collectively exhaustive. Duplication of elements in the sample may result if population elements belonged to more than one cluster. Omissions will result in coverage bias.

Cluster Sampling

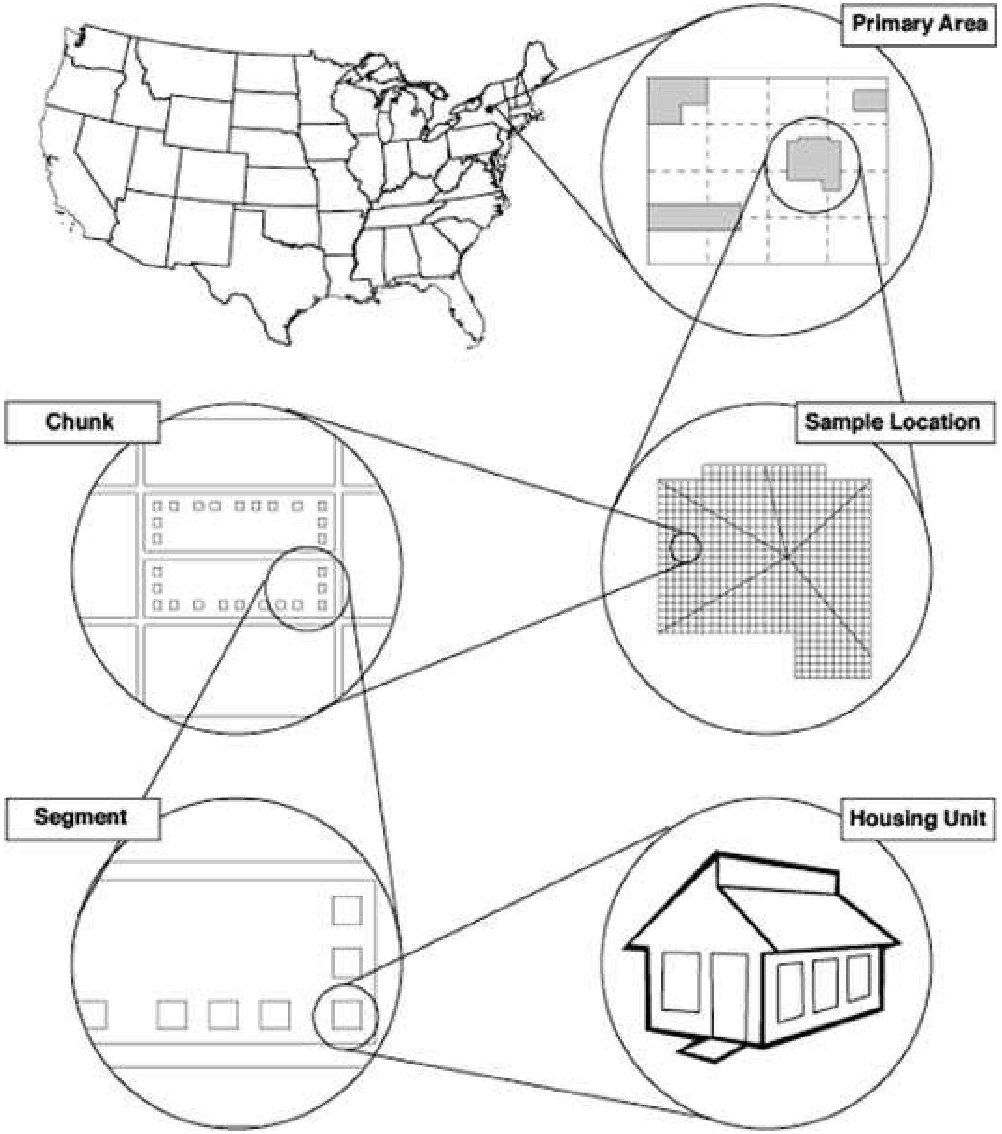
- **Step 5** Determine the number of clusters to be selected. This may be done by dividing the sample size by estimated average number of population elements in each cluster. To the extent the homogeneity and heterogeneity of the clusters are different from that of the population, as cluster number increases, precision increases. On the other hand, as differences between clusters increases, precision decreases.
- **Step 6** Randomly select the targeted number of clusters.

Subtypes of Cluster Sampling

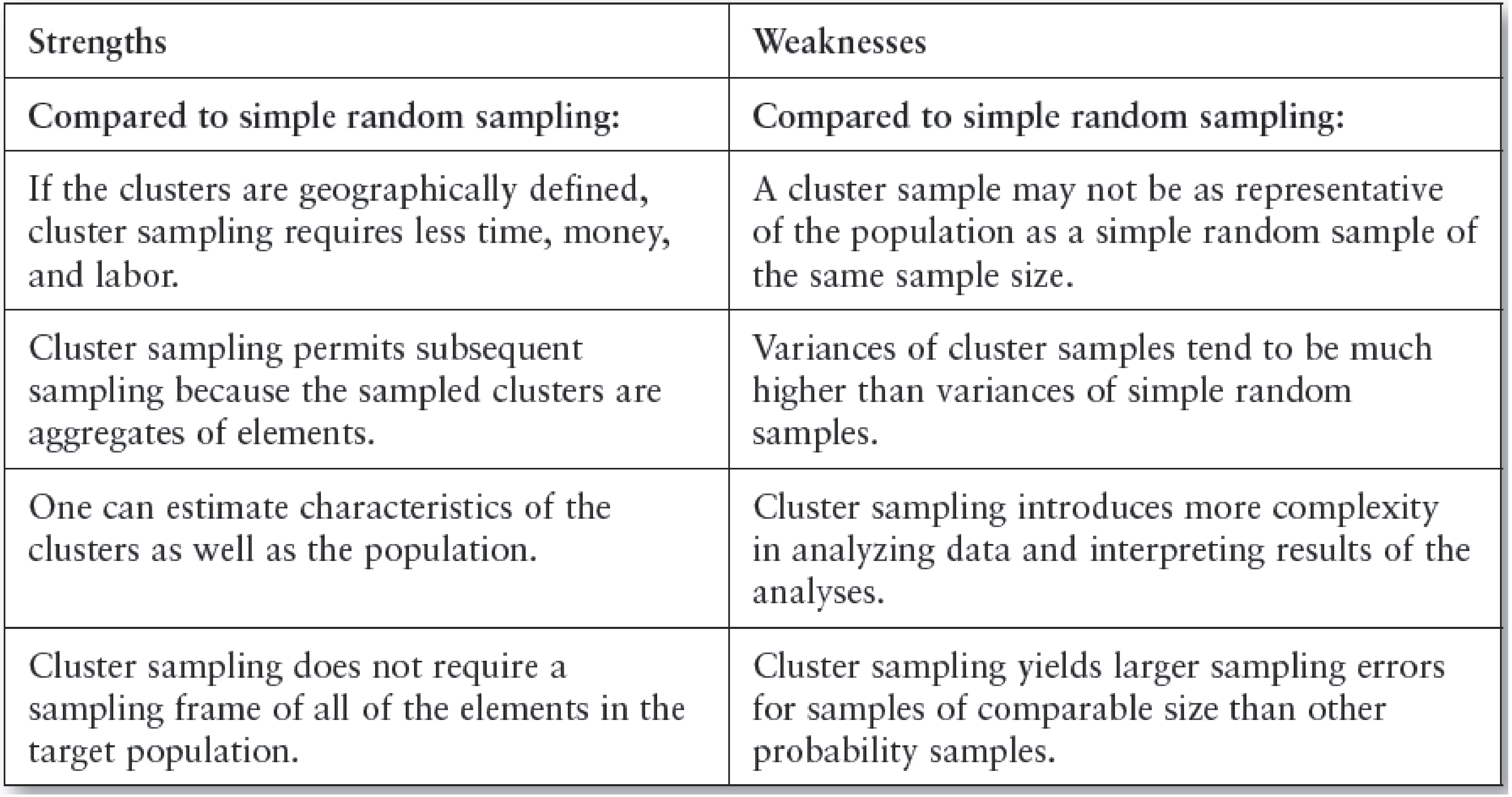
- **Single-stage cluster sampling** In a single-stage cluster sample design, sampling is done only once.

  Example: interest in studying homeless persons who live in shelters. If there are five shelters in a city, a researcher will randomly select one of the shelters and then include in the study all the homeless persons who reside at the selected shelter.
- **Two-stage cluster sampling** A two-stage cluster sample design includes all the steps in single-stage cluster sample design with one exception, the last step. Instead of including all the elements in the selected clusters in the sample, a random sample (either a simple random sample, stratified sample, or systematic sample) is taken from the elements in each selected cluster.
- **Multi-stage cluster sampling** Multistage cluster sampling involves the repetition of two basic steps: listing and sampling. Typically, at each stage, the clusters get progressively smaller in size; and at the last stage element sampling is used. Sampling procedures (simple random sampling, stratified sampling, or systematic sampling) at each stage may differ.

Cluster Sampling

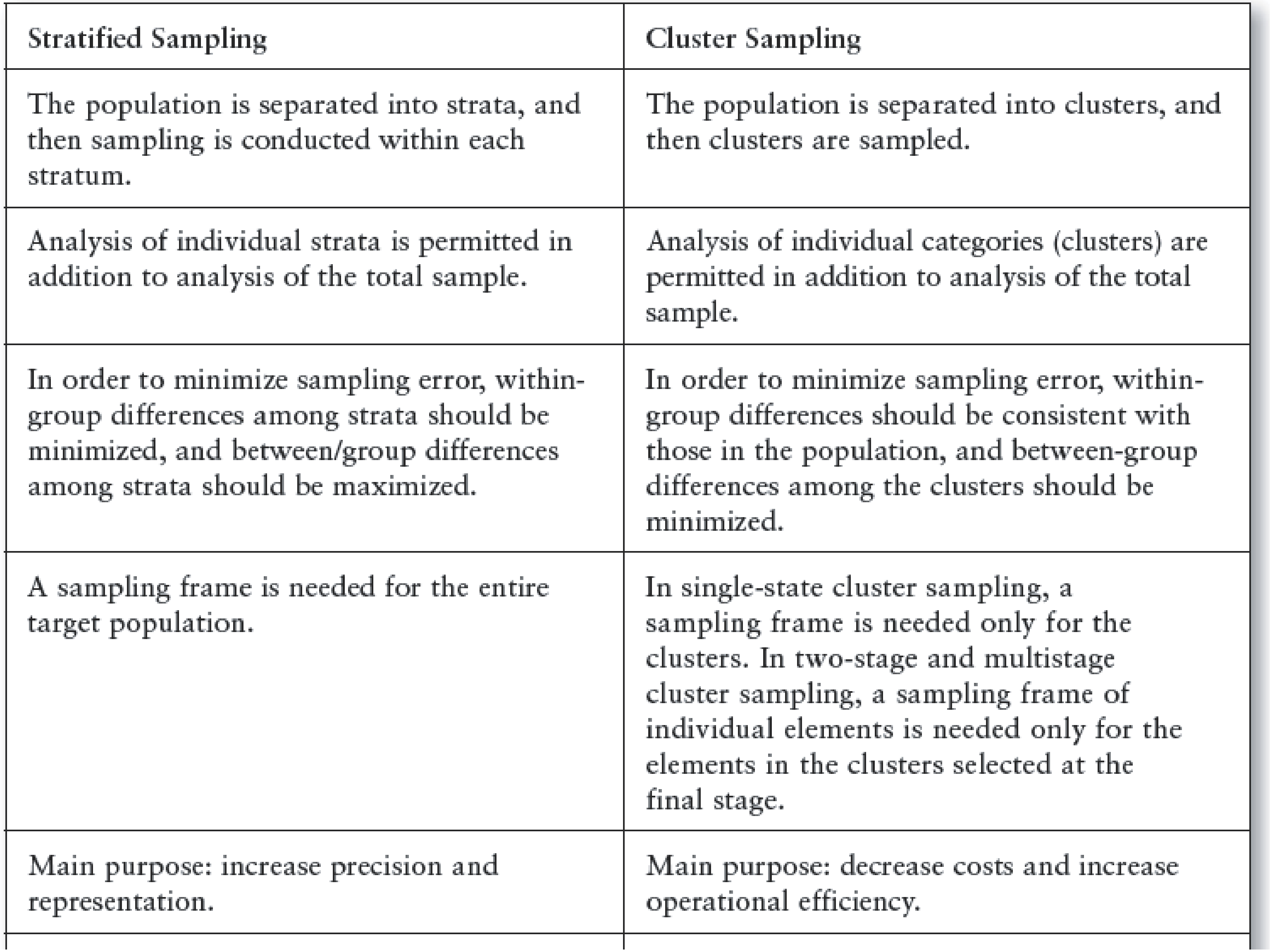


Stenghts and Weaknesses of Cluster Sampling



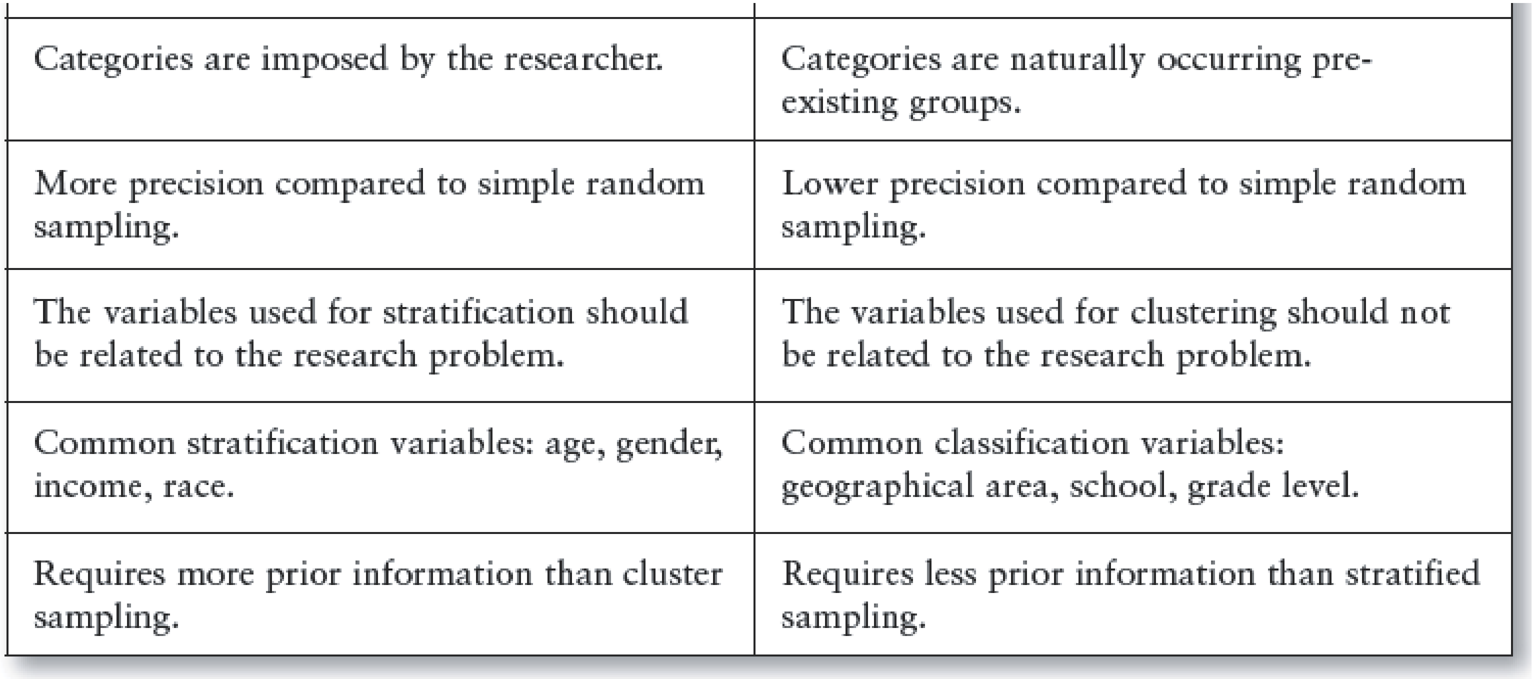
Strengths	Weaknesses
Compared to simple random sampling:	Compared to simple random sampling:
If the clusters are geographically defined, cluster sampling requires less time, money, and labor.	A cluster sample may not be as representative of the population as a simple random sample of the same sample size.
Cluster sampling permits subsequent sampling because the sampled clusters are aggregates of elements.	Variances of cluster samples tend to be much higher than variances of simple random samples.
One can estimate characteristics of the clusters as well as the population.	Cluster sampling introduces more complexity in analyzing data and interpreting results of the analyses.
Cluster sampling does not require a sampling frame of all of the elements in the target population.	Cluster sampling yields larger sampling errors for samples of comparable size than other probability samples.

Comparison Between Cluster Sampling and Stratified Sampling



Stratified Sampling	Cluster Sampling
The population is separated into strata, and then sampling is conducted within each stratum.	The population is separated into clusters, and then clusters are sampled.
Analysis of individual strata is permitted in addition to analysis of the total sample.	Analysis of individual categories (clusters) are permitted in addition to analysis of the total sample.
In order to minimize sampling error, within-group differences among strata should be minimized, and between/group differences among strata should be maximized.	In order to minimize sampling error, within-group differences should be consistent with those in the population, and between-group differences among the clusters should be minimized.
A sampling frame is needed for the entire target population.	In single-state cluster sampling, a sampling frame is needed only for the clusters. In two-stage and multistage cluster sampling, a sampling frame of individual elements is needed only for the elements in the clusters selected at the final stage.
Main purpose: increase precision and representation.	Main purpose: decrease costs and increase operational efficiency.

Comparison Between Cluster Sampling and Stratified Sampling



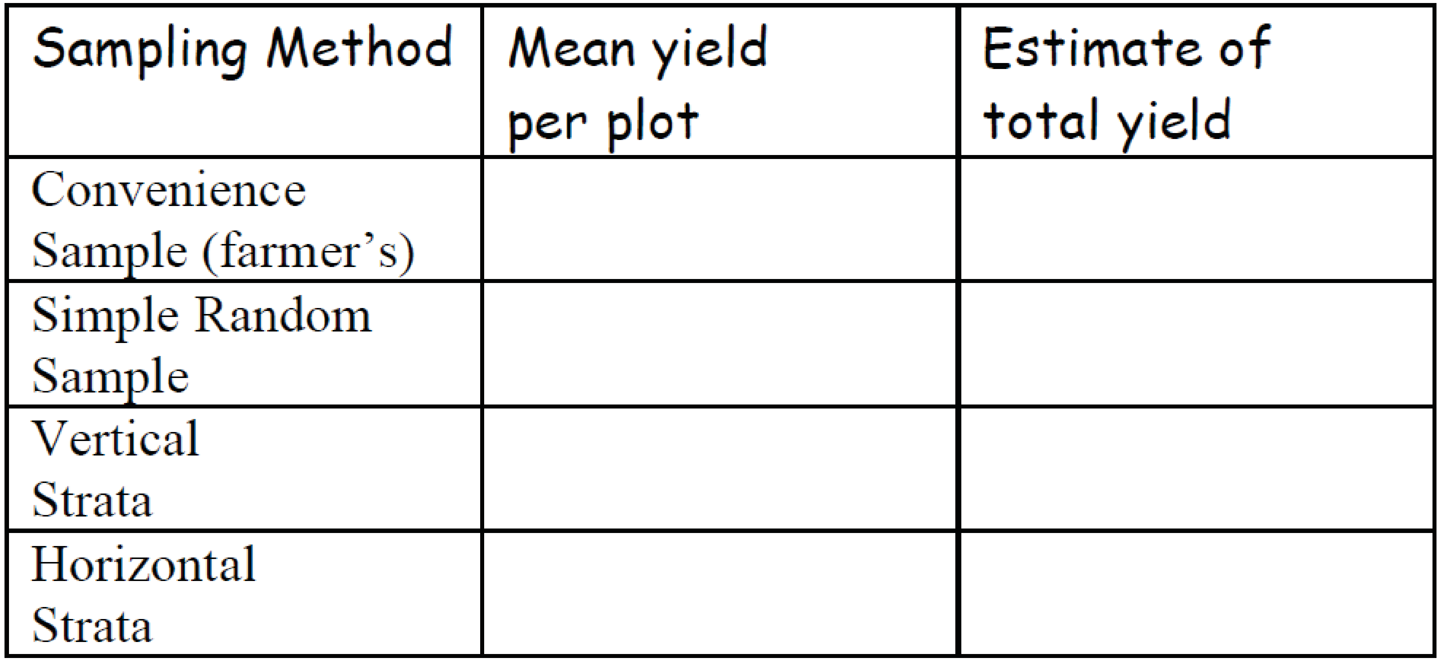
Categories are imposed by the researcher.	Categories are naturally occurring pre-existing groups.
More precision compared to simple random sampling.	Lower precision compared to simple random sampling.
The variables used for stratification should be related to the research problem.	The variables used for clustering should not be related to the research problem.
Common stratification variables: age, gender, income, race.	Common classification variables: geographical area, school, grade level.
Requires more prior information than cluster sampling.	Requires less prior information than stratified sampling.

Rolling Down the River

A farmer has just cleared a new field for corn. It is a unique plot of land in that a river runs along one side. The corn looks good in some areas of the field but not others. The farmer is not sure that harvesting the field is worth the expense. He has decided to harvest 10 plots and use this information to estimate the total yield. Based on this estimate, he will decide whether to harvest the remaining plots.

Rolling Down the River

A farmer has just cleared a new field for corn. It is a unique plot of land in that a river runs along one side. The corn looks good in some areas of the field but not others. The farmer is not sure that harvesting the field is worth the expense. He has decided to harvest 10 plots and use this information to estimate the total yield. Based on this estimate, he will decide whether to harvest the remaining plots.



Sampling Method	Mean yield per plot	Estimate of total yield
Convenience Sample (farmer's)		
Simple Random Sample		
Vertical Strata		
Horizontal Strata		

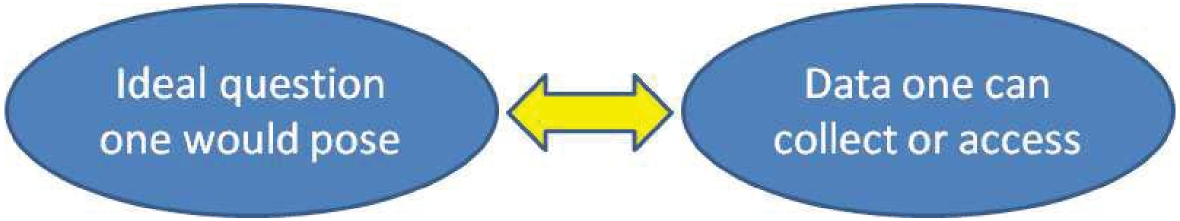
Rolling Down the River

Discussion

- Is there a reason, other than convenience, to choose one method over another?
- How do your estimates vary according to the different sampling methods?
- Do you have similar results as your neighbor?
- What can you tell from comparing "boxplots" of the mean yields under simple random sampling, vertical or horizontal stratified sampling?
- Which sampling method would you promote? Why?
- What is the actual yield of the entire field? How do the boxplots relate to this actual value?

Study Designs: Design Dilemma

A **study design** is a specific plan or protocol for conducting the study, which allows the investigator to translate the conceptual hypothesis into an operational one.



Classification of Study Designs

Qualitative	**Quantitative**
Understanding	Prediction
Interview/observation	Survey/questionnaires
Discovering frameworks	Existing frameworks
Textual (words)	Numerical
Theory generating	Theory testing (experimental)
Quality of informant more important than sample size	Sample size core issue in reliability of data
Subjective	Objective
Embedded knowledge	Public
Models of analysis: fidelity to text or words of interviewees	Model of analysis:parametric, non-parametric

Classification of Study Designs

Qualitative

- Methods
  - Focus Groups
  - Interviews
  - Surveys
  - Self-reports
- Sampling: Purposive
- Quality Assurance:
  - Trustworthiness: e.g., Credibility, Confirmability, Transferability
  - Authenticity: e.g., Educative

Qualitative

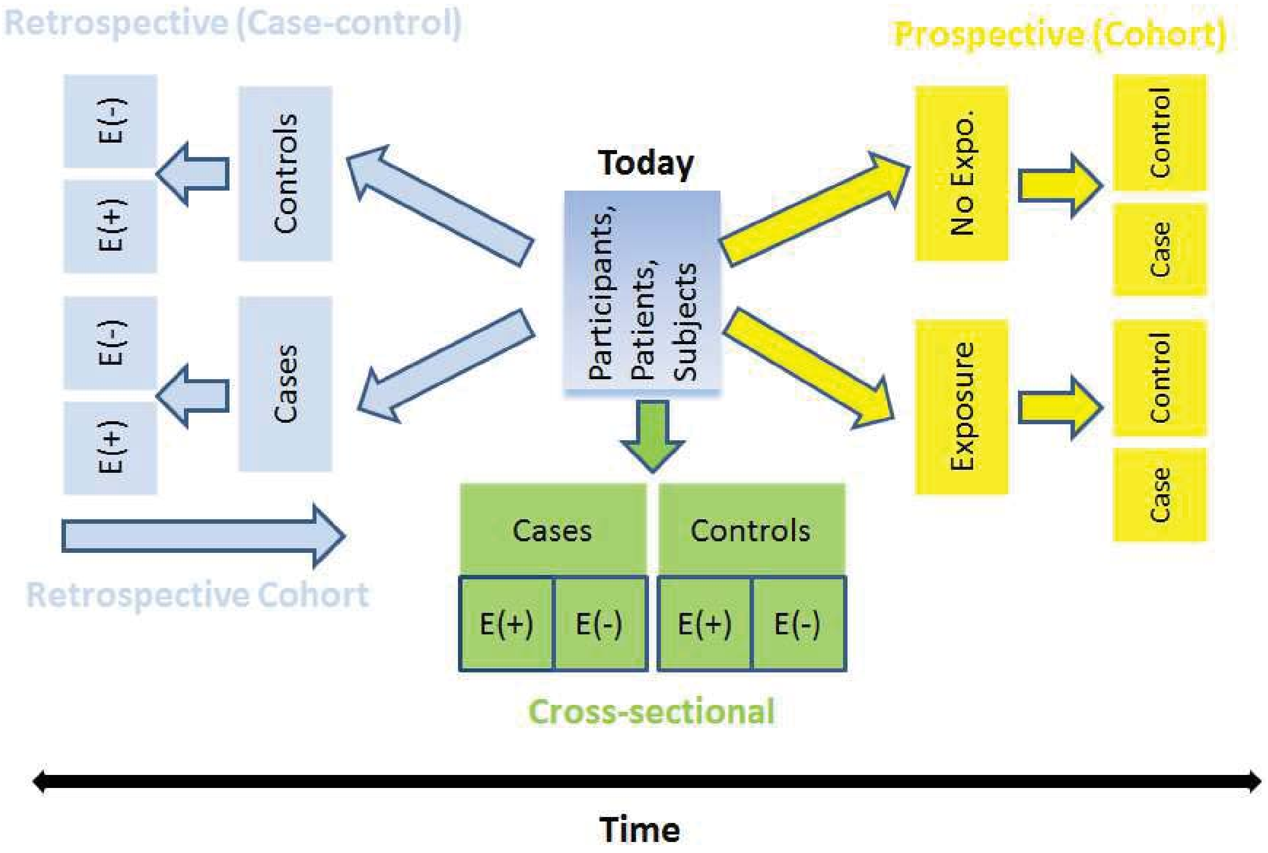
- Methods
  - Observational
  - Experimental
- Sampling: Random (simple, stratified, cluster, etc) or purposive
- Quality Assurance:
  - Reliability: “Consistent”
  - Validity: “Construct”

Reliability and Validity

- **Reliability**:
  - The degree of consistency between two measures of the same thing. (Mehrens and Lehman, 1987).
  - The measure of how stable, dependable, trustworthy, and consistent a test is in measuring the same thing each time (Worthen et al., 1993)
- **Validity**:
  - Truthfulness: Does the test measure what it purports to measure? the extent to which certain inferences can be made from test scores or other measurement. (Mehrens and Lehman, 1987)
  - The degree to which they accomplish the purpose for which they are being used. (Worthen et al., 1993)

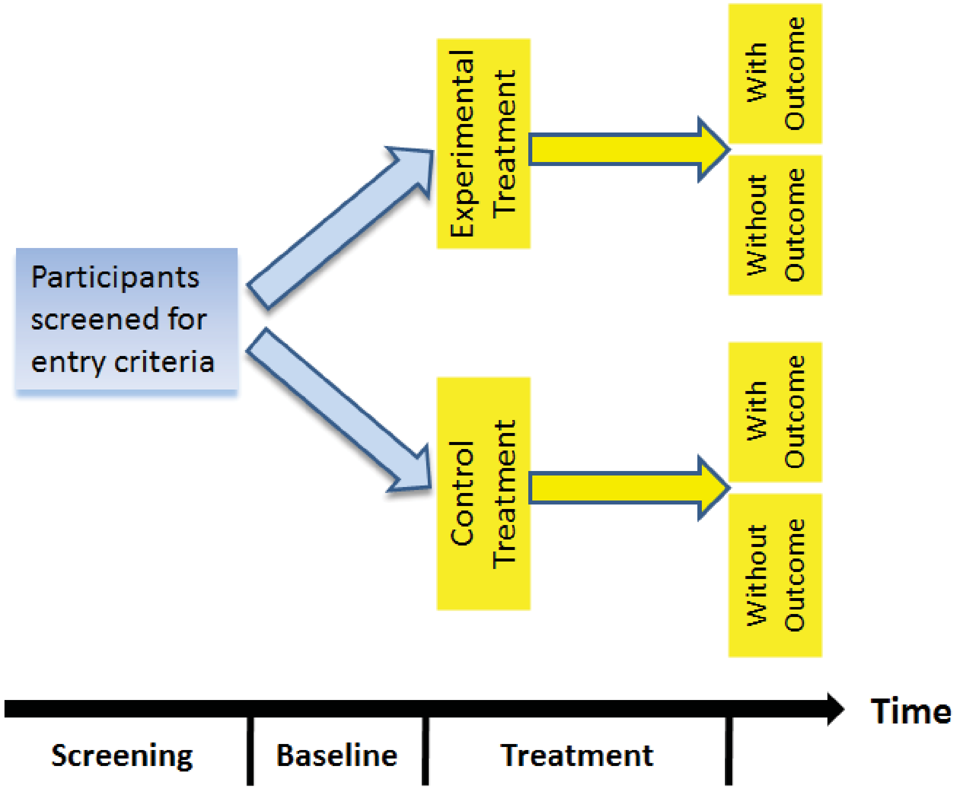
Observational Study Designs

Observational: studies that do not involve any intervention or experiment.

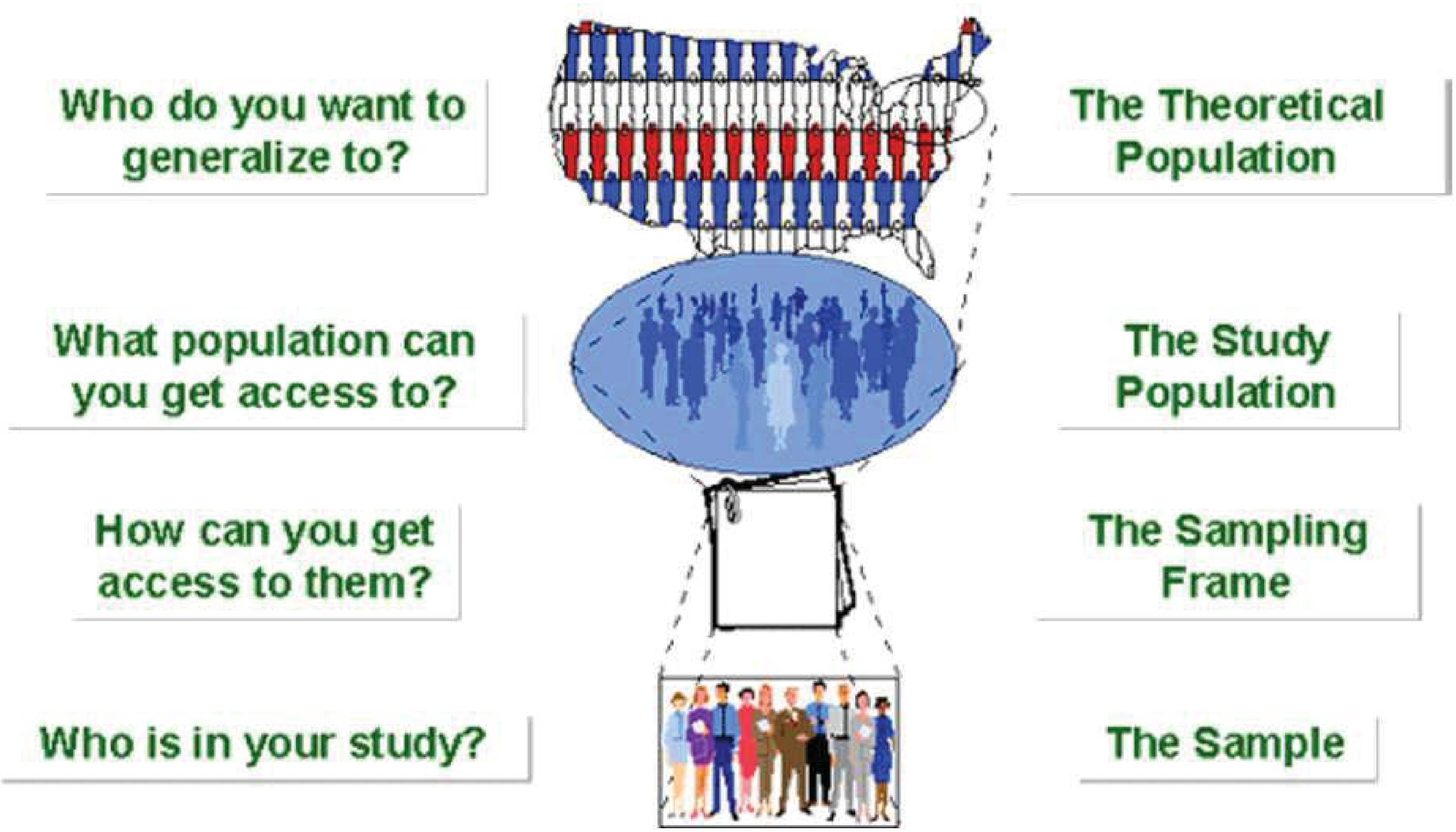


Experimental Study Designs

Experimental: studies that entail manipulation of the study factor (exposure) and randomization of subjects to treatment (exposure) groups.



Popular Statistics and their Distribution



Who do you want to generalize to? — The Theoretical Population

What population can you get access to? — The Study Population

How can you get access to them? — The Sampling Frame

Who is in your study? — The Sample

Some Definitions

- **Distribution of sample** Let $X_1, X_2, \ldots, X_n$ denote a sample of size n. The distribution of the sample $X_1, X_2, \ldots, X_n$ is defined to be the joint distribution of $X_1, X_2, \ldots, X_n$.
- Hence, if $X_1, X_2, \ldots, X_n$ is a random sample of size n from $f(.)$ then the distribution of the random sample $X_1, X_2, \ldots, X_n$, defined as the joint distribution of $X_1, X_2, \ldots, X_n$, is given by $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$, and $X_1, X_2, \ldots, X_n$ are stochastically independent.
- **Statistic** Any function of the elements of a random sample, which does not depend on unknown parameters, is called a statistic.
- Statistics may serve as estimators for a parameter of interest.

Examples of Statistics

- $\overline{X} = \sum_{i=1}^{n} X_i/n$ is called the **sample mean**.
- $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ is called the **sample variance** (sometimes denoted as S_{n-1}^2).
- $S = \sqrt{S^2}$ is called the **sample standard deviation**.
- $M_r = \sum_{i=1}^{n} X_i^r/n$ is called the r**th sample moment** about the origin.
- Suppose that the random variables $X_1, \ldots, X_n$ are ordered and re-written as $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The vector $(X_{(1)}, \ldots, X_{(n)})$ is called the **ordered sample**.
- The **standard error** (SE) is an estimate of the standard deviation of a statistic. It is important because it is used to compute other measures, like confidence intervals or margins of error (see later)

Sampling Distributions

- When the distribution of interest consists of *all* the unique samples of size n that can be drawn from a population, the resulting distribution of sample means is called the **sampling distribution of the mean**.
- We can generate a distribution of anything, as long as we have values / scores to work with (cfr tossing a coin and scoring the sample wrt nr of heads).
- There are also sampling distributions of medians, standard deviations, and any other statistic you can think of.
- In other words, populations, which are distributions of individual elements, give rise to sampling distributions, which describe how collections of elements are distributed in the population.

Sampling Distributions

Level	**Collection**	**Elements**
Population	All individuals ($N =$ size of population)	The scores each individual receives on some attribute
Sample	Subset of individuals from the population ($n =$size of sample)	The scores each individual in the sample receives on some attribute
Sampling Distribution	All unique samples of size n from the population	The values of a statistic applied to each sample

Sampling Distributions

- In inferential statistics we make use of two important properties of sampling distributions, expressed in lay terms as:
    - The mean of all unique samples of size n (i.e., the average of all the means) is identical to the mean of the population from which those samples are drawn. Thus, any claims about the mean of the sampling distribution apply to the population mean.
    - The shape of the sampling distribution increasingly approximates a normal curve as sample size n is increased, even if the original population is not normally distributed.
- If the original population is itself normally distributed, then the sampling distribution will be normally distributed even when the sample size is only one.

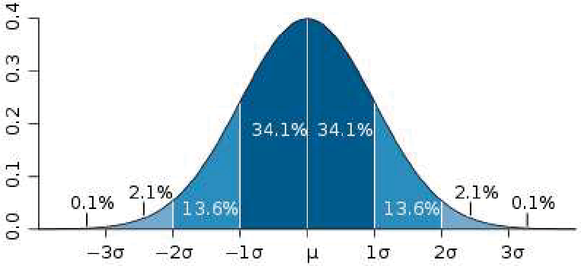
The Empirical Rule

- Sometimes the door handles in office buildings show a wear pattern caused by thousands, maybe millions of times being pulled or pushed to open the door. Often you will see that there is a middle region that shows by far the most amount of wear at the place where people opening the door are the most likely to grab the handle, surrounded by areas on either side showing less wear. On average, people are more likely to have grabbed the handle in the same spot and less likely to use the extremes on either side.

- Many real-life phenomena are “normal”.

The Empirical Rule

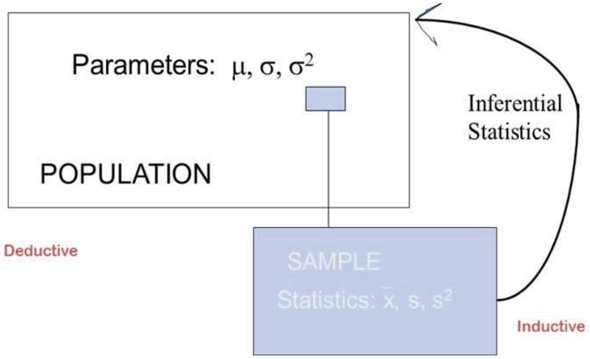
- The so-called **empirical rule** states that the bulk of a set of data will cluster around the mean in the following fashion:
  - 68% of values fall within 1 standard deviation of the mean
  - 95% fall within ±2 standard deviations of the mean
  - 99% fall within ±3 standard deviations of the mean
- It is called the "empirical rule" since experimenters have observed roughly these patterns from their data over and over again when they empirically collect data.



Theoretical Sampling Distributions

- Unless the details of a population are known in advance, it is not possible to perfectly describe any of its sampling distributions.
  - When the population details are known, we can simply calculate the desired parameter, and then there would be no point in collecting samples.
- For this reason, a variety of idealized, theoretical sampling distributions have been described mathematically, including the student t distribution or the F distribution (see later: theory and practical sessions), which can be used as statistical models for the real sampling distributions.
- The theoretical sampling distributions can then be used to obtain the likelihood (or probability) of sampling a particular mean if the mean of the sampling distribution (and hence the mean of the original population) is some particular value. The population parameter will first have to be hypothesized, as the true state of affairs is generally unknown. This is called the null hypothesis (see later: Chapter "Hypothesis Testing")

Computing the Standard Error: a Measure of Sampling Error


Parameters: μ, σ, σ^2
POPULATION
Inferential Statistics
Deductive
SAMPLE
Statistics: $\bar{x}$, s, s^2
Inductive

Population parameter	**Sample statistic**
N: Number of observations in the population N_i: Number of observations in population i P: Proportion of successes in population P_i: Proportion of successes in population i μ: Population mean μ_i: Mean of population i σ: Population standard deviation σ_p: Standard deviation of p $\sigma_{\overline{x}}$: Standard deviation of $\overline{x}$	n: Number of observations in the sample n_i: Number of observations in sample i p: Proportion of successes in sample p_i: Proportion of successes in sample i $\overline{x}$: Sample estimate of population mean x_i: Sample estimate of μ_i s or S: Sample estimate of σ SE_p: Standard error of p $SE_{\overline{x}}$: Standard error of x

Computing the Standard Error: a Measure of Sampling Error

- The variability of a statistic is measured by its standard deviation. The table below show formulas for computing the standard deviation of statistics from simple random samples.

Statistic	**Standard Deviation**
Sample mean, $\overline{x}$	$\sigma_{\overline{x}} = \sigma/\sqrt{n}$
Sample proportion, p	$\sigma_p = \sqrt{P(1-P)/n}$
Difference between means, $\overline{x_1} - \overline{x_2}$	$\sigma_{\overline{x_1} - \overline{x_2}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
Difference between proportions, $p_1 - p_2$	$\sigma_{p_1 - p_2} = \sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}$

- So in order to compute the standard deviation of a sample statistic (spread in sample distribution of the statistic), you must know the value of one or more population paramaters.

Computing the Standard Error: a Measure of Sampling Error

- The values of population parameters are often unknown, making it impossible to compute the standard deviation of a statistic. When this occurs, use the standard error.
- The standard error is computed from known sample statistics, and it provides an unbiased estimate of the standard deviation of the statistic. The table below shows how to compute the standard error for simple random samples, assuming the population size is at least 10 times larger than the sample size.

Statistic	**Standard Error**
Sample mean, $\overline{x}$	$SE_{\overline{x}} = s/\sqrt{n}$
Sample proportion, p	$SE_p = \sqrt{p(1-p)/n}$
Difference between means, $\overline{x_1} - \overline{x_2}$	$SE_{\overline{x_1}-\overline{x_2}} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$
Difference between proportions, $p_1 - p_2$	$SE_{p_1-p_2} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$

Taylor Expansion - An aside to compute variances

- **Definition**: If a function $g(x)$ has derivatives of order r, that is $g^{(r)}(x) = \frac{\partial^r}{\partial x^r} g(x)$ exists, then for any constant a, the Taylor expansion of order r about a is:

$$T_r(x) = \sum_{k=0}^{r} \frac{g^{(k)}(a)}{k!}(x-a)^k.$$

- The major theorem from Taylor is that the remainder from the approximation, namely $g(x) - T_r(x)$, tends to 0 faster than the highest-order term in $T_r(x)$.
- **Theorem**: If $g^{(r)}(a) = \frac{\partial^r}{\partial x^r} g(x)|_{x=a}$ exists, then

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

- For the purpose of the Delta Method, we will only use $r = 1$.

Applying the Taylor Theorem

- Let $T_1, T_2, \ldots, T_k$ be random variables with means $\theta_1, \theta_2, \ldots, \theta_k$ and define $T = (T_1, \ldots, T_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$.
- The first order Taylor series expansion of g about θ (multivariate version) is

$$g(t) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i) + \text{Remainder}$$

- Taking expectations from both sides:

$$E[g(T)] \approx g(\theta)$$

- We can also approximate the variance of $g(T)$ by

$$\begin{aligned} Var[g(T)] &\approx E[(g(T) - g(\theta))^2] \\ &\approx E[(\sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i))^2] \\ &= \sum_{i=1}^{k} g_i'(\theta)^2 Var(T_i) + 2\sum_{i>j} g_i'(\theta) g_j'(\theta) Cov(T_i, T_j) \end{aligned}$$

Replicated sampling

- The essence of this strategy is to facilitate the variance calculation by selecting a set of replicated subsamples instead of a single sample.
- It requires each subsample to be drawn independently and to use an identical sample selection design.
- Then an estimate is made in each subsample by the identical process, and the sampling variance of the overall estimate (based on all subsamples) can be estimated from the variability of these independent subsample estimates.

Jacknife and bootstrap

- The **jackknife procedure** is to estimate the parameter of interest n times, each time deleting one sample data point. The average of the resulting estimators, called "pseudovalues", is the jackknife estimate for the parameter. For large n, the jackknife estimate is approximately normally distributed about the true parameter.
- The **bootstrap method** involves drawing samples repeatedly from the empirical distribution. So in practice, a large number of samples of size n are drawn with replacement, from the original n data points. Each time, the parameter of interest is estimated from the bootstrap sample, and the average over all bootstrap samples is taken to be the bootstrap estimate of the parameter of interest.

In summary

How can we know whether our sample is representative of the underlying population?

- Avoid small samples, as there are more extreme (i.e., rare) sample means in the sampling distribution, and we are more likely to get one of them in an experiment.
- We have control over sampling error because sample size determines the standard error (variability) in a sampling distribution.
- We will see that sample size is closely connected to the concept of "power": if a specific power is targeted to identify an effect in a testing strategy, then one can compute the necessary sample size to achieve the pre-specified power of the test.
- On a practical note:
  - Realize that large samples are not always attainable and that clever more complicated sample strategies than simple random sampling need to be followed.
  - A correction is needed when sampling from a finite distribution

A note aside

- The central limit theorem and the formulae for standard errors of the mean and the proportion are based on the premise that the samples selected are chosen with replacement.
- However, in virtually all survey research, sampling is conducted without replacement from populations that are of a finite size n_p
- In these cases, particularly when the sample size n is not small in comparison with the population size n_p (i.e., more than 5% of the population is sampled) so that $n > n_p 0.05$, a finite population correction factor (fpc) is used to define for instance both the standard error of the mean and the standard error of the proportion.

A note aside

- If we denote the mean and standard deviation of the sampling distribution of means by $\mu_{\overline{x}}$ and $\sigma_{\overline{x}}$, and the population mean and standard deviation by μ and σ, then actually

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{n_p - n}{n_p - 1}}.$$

- If the population is infinite or if sampling is with replacement, the above result reduces to

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}},$$

  in line with "large sample theory" and the Central Limit Theorem.

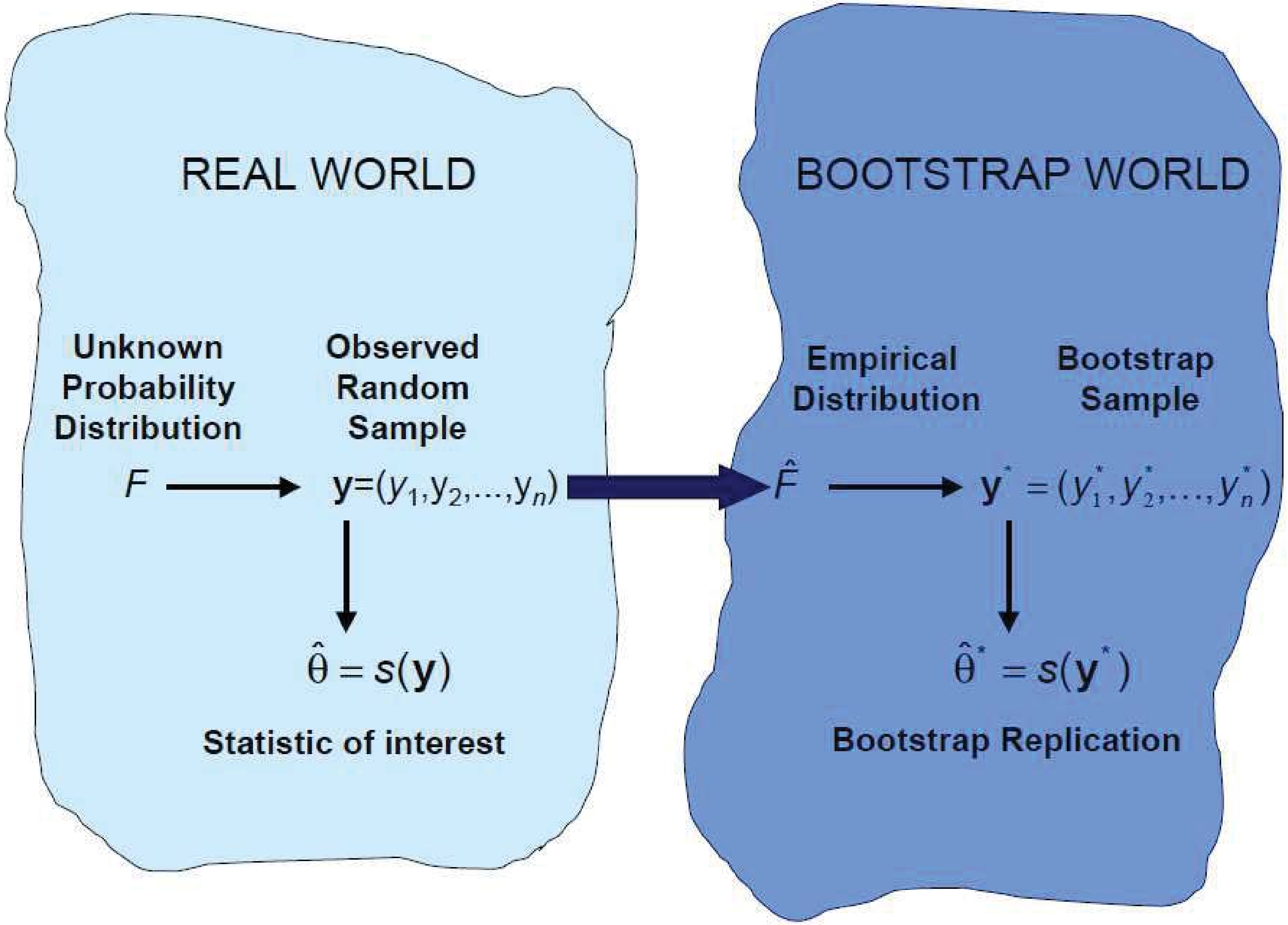
Resampling

- Approximations obtained by random sampling or simulation are called **Monte Carlo** estimates.
- Assume random variable Y has a certain distribution. Use simulation or analytic derivations to study how an estimator, computed from samples from this distribution, behaves: e.g., Y has lognormal distribution, what is the variance of the median?
  - Analytical solution? Need knowledge of the population distribution function
  - Simulate 500 samples of size n from the lognormal distribution, compute the sample median for each sample, and then compute the sample variance of the 500 sample medians.

The Bootstrap

- Efrons bootstrap is a general purpose technique for obtaining estimates of properties of statistical estimators without making assumptions about the distribution of the data.
- Often used to find:
  - Standard errors of estimates (may be easier than "Delta Method")
  - Confidence intervals for unknown paramters (see later "Confidence Intervals")
  - *p*-values for test statistics under a null hypothesis (see later "Hypothesis testing")

The Bootstrap



REAL WORLD

Unknown Probability Distribution

Observed Random Sample

$F \longrightarrow \mathbf{y}=(y_1,y_2,\ldots,y_n)$

$\hat{\theta} = s(\mathbf{y})$

Statistic of interest

BOOTSTRAP WORLD

Empirical Distribution

Bootstrap Sample

$\hat{F} \longrightarrow \mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$

$\hat{\theta}^* = s(\mathbf{y}^*)$

Bootstrap Replication

The Bootstrap

- Suppose Y has a cumulative distribution function (cdf) $F(y) = P(Y \leq y)$.
- We have a sample of size n from $F(y)$, $Y_1, Y_2, \ldots, Y_n$
- **Steps**:
  - Repeatedly simulate sample of size n from F
  - Compute statistic of interest
  - Study behavior of statistic over B repetitions

The Bootstrap

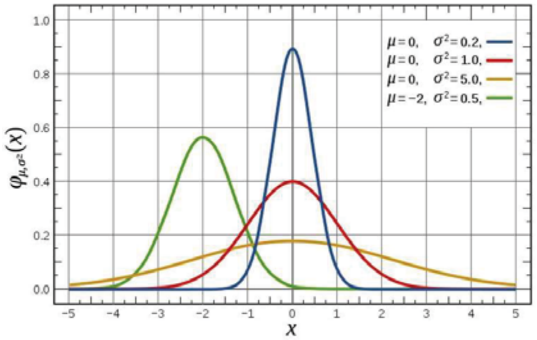
- Without knowledge of F use the empirical cdf $F_n(y) = 1/n \sum_{i=1}^{n} I(Y_i \leq y)$ as estimate of F.
- Pretend that $F_n(y)$ is the original distribution $F(y)$.
- Sampling from $F_n(y)$ is equivalent to sampling with replacement from originally observed $Y_1, \ldots, Y_n$.
- Special case: leave-one-out observation samples $=$ **Jackknife** samples

Inverse Transform Method for Simulating Continuous Random Variables

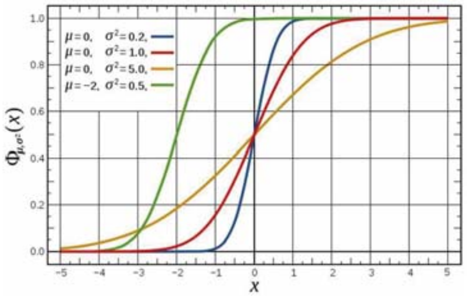
- The **Method of Inverse Transforms** is most often used to simulate a realization of a random variable associated with a particular distribution. Inverse transform sampling works as follows.
- Consider, for example, a continuous random variable with cumulative distribution function F.
- Let U be a uniform random variable over the unit interval and pass U through the inverse of the cumulative distribution function, that is, compute $X = F^{-1}(U)$, where X constitutes a "sample".
- It can be seen that for a sufficiently large set of samples, the associated normalized histogram generates a close approximation to the probability density function of the random variable associated with the cumulative distribution function F.

Inverse Transform Method for Simulating Continuous Random Variables

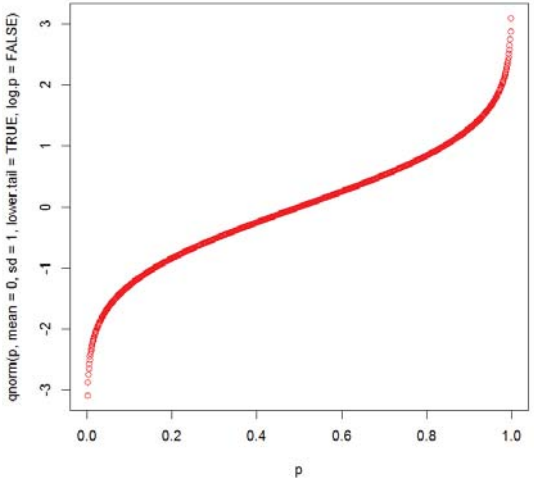
Standard Normal



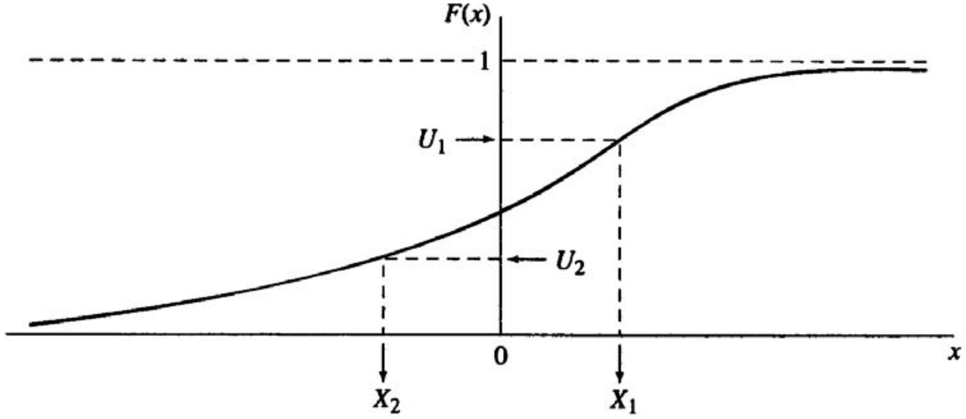
Cumulative Standard Normal



Inverse Cumulative Standard Normal



Idea behind Quantile Function

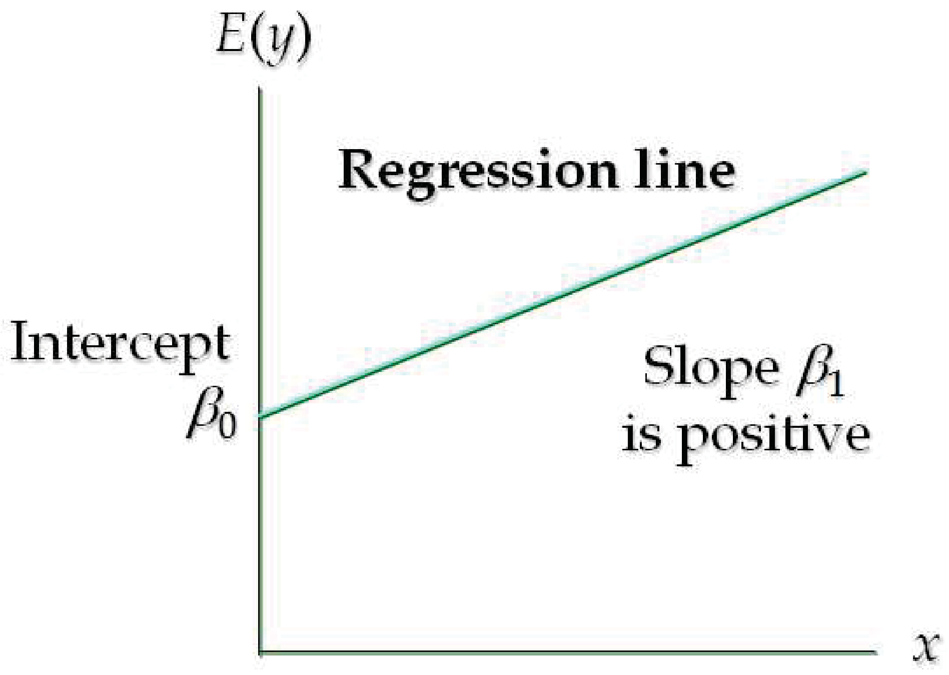


Exploratory Data Analysis: Motivating example

- Given 4 data sets (actual data omitted), for which
  - N = 11
  - Mean of X = 9.0
  - Mean of Y = 7.5
  - Intercept (β_0) = 3
  - Slope (β_1) = 0.5
  - Residual standard deviation = 1.236
  - Correlation = 0.816 (0.817 for data set 4)
- $Y = \beta_0 + \beta_1 X + \epsilon$, with ϵ a random variable called the error-term, and β_0, β_1 parameters, is called a **simple linear regression model**. In such a model, it is assumed that the expecation of Y given X is $E(Y) = \beta_0 + \beta_1 X$ (see later).

Do you think these 4 data sets will give equivalent results?

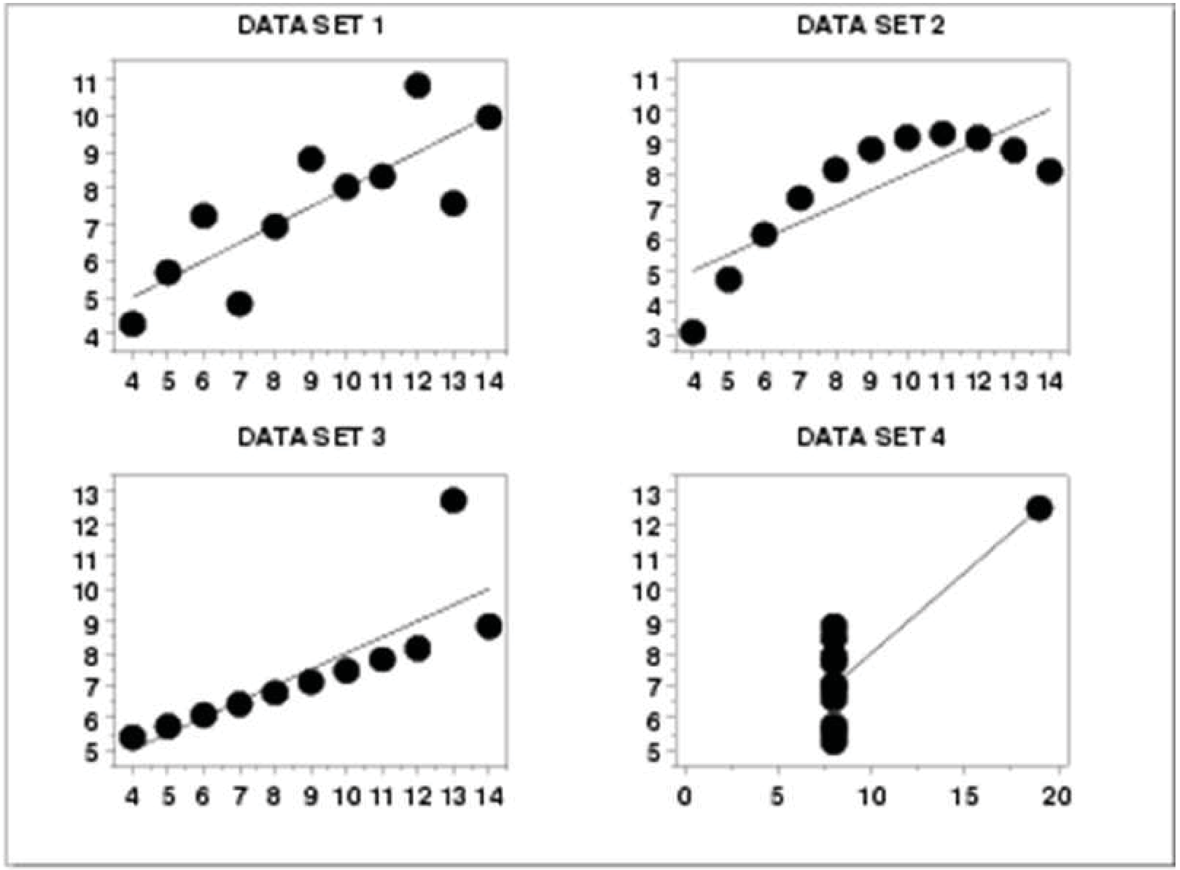
Motivating example



Do you think the four aforementioned data sets will give equivalent results?

Motivating example

- A "scatter plot" of each data set (i.e., plotting Y values versus corresponding X values in a plane), would be the first step of any EDA approach . . . and would immediately reveal non-equivalence!



Data analysis procedures

- There are three popular data analysis approaches

classical analysis	Problem → Data → Model → Analysis → Conclusions
Bayesian analysis	Problem → Data → Model → Prior Distribution → Analysis → Conclusions
EDA	Problem → Data → Analysis → Model → Conclusions

Data analysis procedures

- For **classical analysis**, the data collection is followed by proposing a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.
- For a **Bayesian analysis**, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model: the analysis formally combines both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.
- For **EDA**, the data collection is not followed by a model imposition: it is followed immediately by analysis with a goal of inferring what model would be appropriate

Data analysis procedures

- Moreover, statistics and data analysis procedures can broadly be split into two parts:
  - Quantitative procedures
  - Graphical procedures
- **Quantitative techniques** are the set of statistical procedures that yield numeric or tabular output:
  - hypothesis testing
  - analysis of variance (is there more variation within groups of observations than between groups of observations?)
  - point estimates and confidence intervals
  - least squares regression
- These and similar techniques are all valuable and are mainstream in terms of classical analysis.

Data analysis procedures

- The **graphical techniques** are for a large part employed in an Exploratory Data Analysis framework. They are often quite simple:
  - plotting the raw data such as via histograms, probability plots
  - plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
  - positioning such plots so as to maximize our natural pattern-recognition abilities (multiple plots, when grouped together, may give a more complete picture of what is going on in the data)

Exploratory Data Analysis

- EDA is not identical to statistical graphics (although the two terms are used almost interchangeably) . . . It is much more.
- Statistical graphics is a collection of graphically-based techniques. They are all focusing on data characterization aspects.
- EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

Exploratory Data Analysis is an approach/philosophy for data analysis that employs a variety of techniques. The main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so.

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to
  - maximize insight into a data set;
  - uncover underlying structure;
  - extract important variables;
  - test underlying assumptions;
  - develop parsimonious models;
  - detect outliers and anomalies

Outlier Detection

- Definition of Hawkins [Hawkins 1980]:

  An **outlier** is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism
- Statistics-based intuition
  - "Normal data" objects follow a "generating mechanisms", e.g. some given statistical process
  - "Abnormal objects" deviate from this generating mechanism
- Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Should one always discard outlying observations?

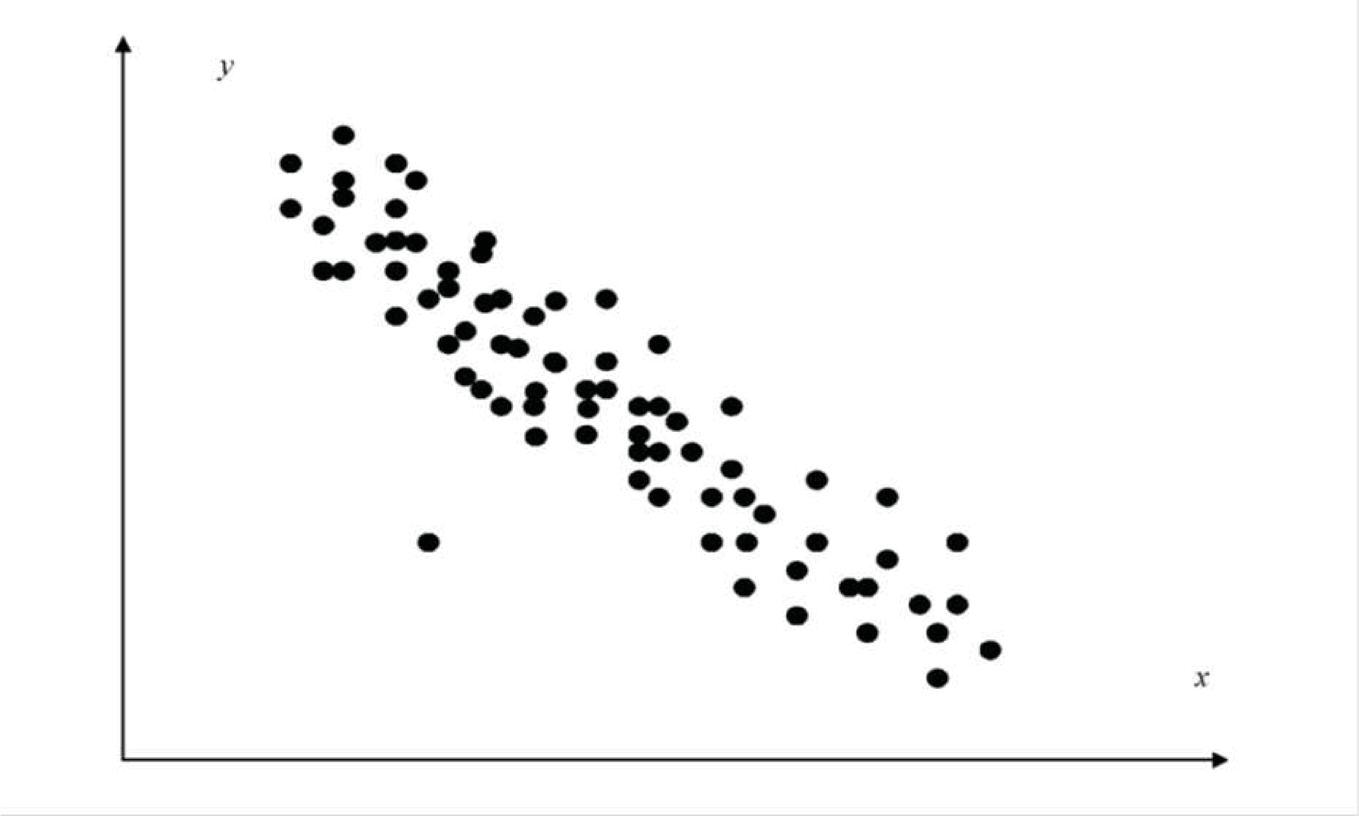
Examples (I)

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age,)
- Public health
  - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city

Examples (II)

- Sports statistics
  - In many sports, various parameters are recorded for players in order to evaluate the players' performances
  - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
  - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
- Detecting measurement errors
  - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
  - Abnormal values could provide an indication of a measurement error
  - Removing such errors can be important in other data mining and data analysis tasks

Univariate and Multivariate Views



Food for Thought

- Data usually are multivariate, i.e., multi-dimensional: The basic model for outliers is univariate, i.e. one-dimensional
- There is usually more than one generating mechanism/statistical process underlying the "normal" data: basic model assumes only one "normal" generating mechanism
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers: The basic model assumes that outliers are rare observations
- A lot of models and approaches have evolved in the past years in order to extend these assumptions: For instance, extreme-value analysis techniques
- Remember: One person's noise could be another person's signal!

Methods of EDA: one-way

- Ordering : Stem-and-Leaf plots
- Grouping: frequency displays, distributions; histograms
- Summaries: summary statistics, standard deviation, box-and-whisker plots

Stem-and-leaf Plots

Age in years (10 observations):

25, 26, 29, 32, 35, 36, 38, 44, 49, 51

Age Interval	Observations
20-29	5 6 9
30-39	2 5 6 8
40-49	4 9
50-59	1

Stem-and-leaf Plots

- The age interval is the **stem**
- The observations are the **leaves**
- Rule of thumb:
  - The number of stems should roughly equal the square root of the number of observations
  - Or the stems should be logical categories

Cumulative Frequency Distribution Tables

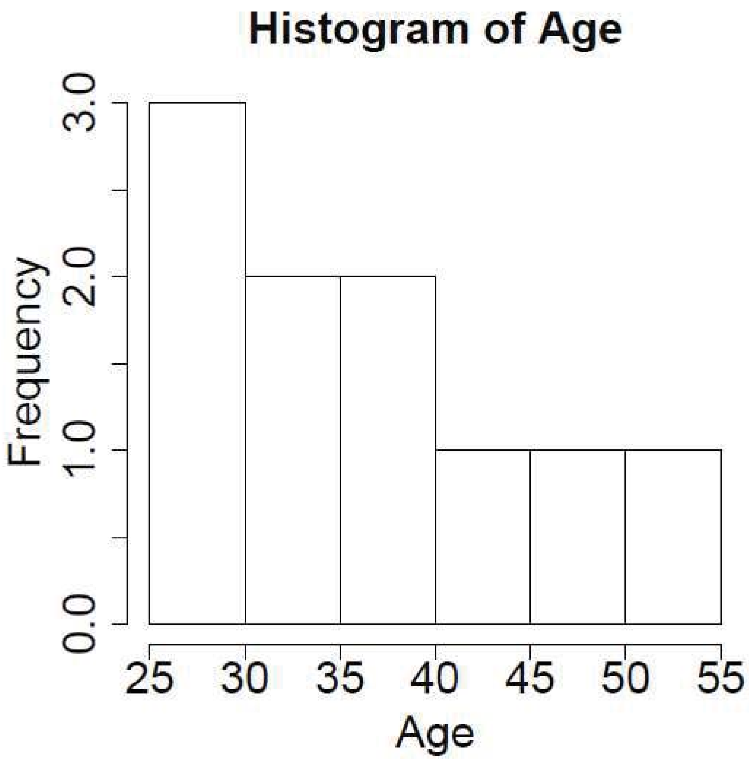
Show the frequency, the relative frequency, and cumulative frequency of observations

Age Interval	Frequency	Cum. Freq.	Rel. Freq	Cum. Rel. Freq.
20-29	3	3	0.3	0.3
30-39	4	7	0.4	0.7
40-49	2	9	0.2	0.9
50-59	1	10	0.1	1.0

- This table shows an **empirical distribution function** obtained from a sample
- The true distribution function is the distribution of the entire population

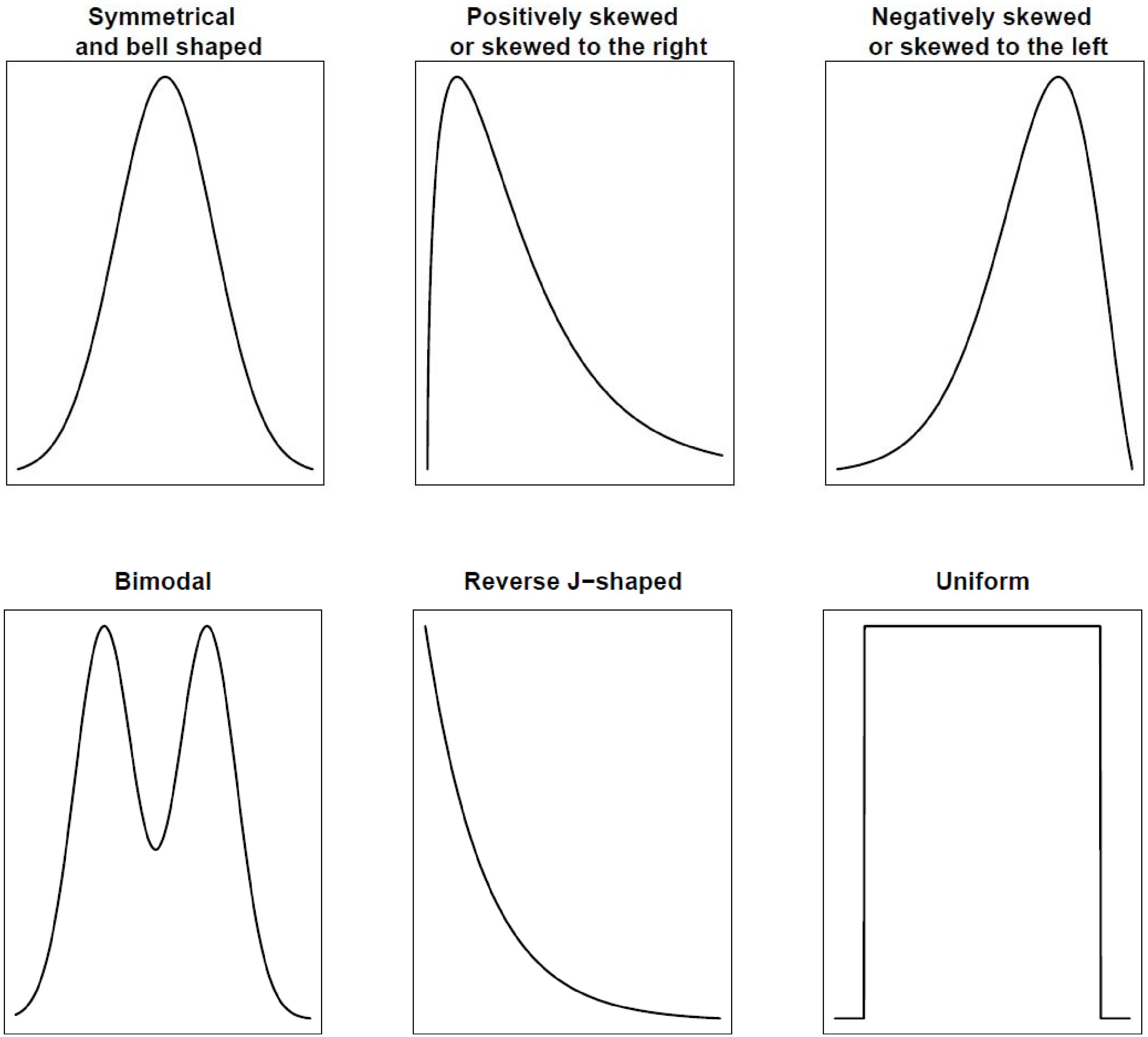
Histograms

Picture of the frequency or relative frequency distribution



Note: Graphs are generally better to use in presentations than tables. They allow your audience to visualize a trend quickly.

Histograms and densities



Symmetrical and bell shaped

Positively skewed or skewed to the right

Negatively skewed or skewed to the left

Bimodal

Reverse J−shaped

Uniform

Summary statistics: percentiles

- The **r-th percentile** P_r of a set of values, divides them such that r percent of the values lie below and $(100 - r)$ percent of the values above.

Percentile	Quartile	Formula
P_{25}	Q_1	${\frac{n+1}{4}}^{th}$ observation
P_{50}	Q_2	${\frac{n+1}{2}}^{th}$ observation
P_{75}	Q_3	${\frac{3(n+1)}{4}}^{th}$ observation

Calculating quartiles

From the age data:

25, 26, 29, 32, 35, 36, 38, 44, 49, 51

with n=10

$$\begin{aligned} Q_2 &= \text{median} \\ &= \text{average of } 5^{th} \text{ and } 6^{th}\text{observations} \\ &= \frac{35+36}{2} \\ &= 35.5 \end{aligned}$$

Remember to order your data!

Calculating quartiles

$$\begin{aligned} Q_1 &= \text{median of lower half of data} \\ &= \text{third smallest value} \\ &= 29 \end{aligned}$$

$$\begin{aligned} Q_3 &= \text{median of upper half of data} \\ &= \text{third largest value} \\ &= 44 \end{aligned}$$

Note: If n is odd, include the median in the upper and lower half of the data.

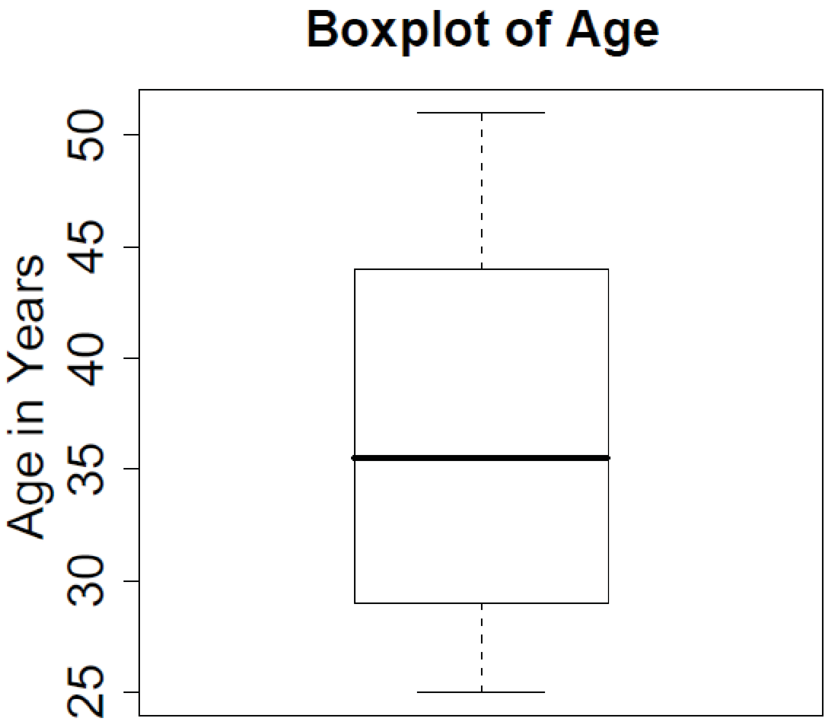
Box-and-whisker plots

- Box-and-whisker plots display quartiles
- Some terminology:
  - Upper Hinge $= Q_3 =$ Third quartile
  - Lower Hinge $= Q_1 =$ First quartile
  - Interquartile range (IQR) $= Q_3 - Q_1$

    Contains the middle 50% of data
  - Upper Fence $=$ Upper Hinge $+$ 1.5 * (IQR)
  - Lower Fence $=$ Lower Hinge - 1.5 * (IQR)
  - Outliers: Data values beyond the fences
- Whiskers are drawn to the smallest and largest observations within the fences

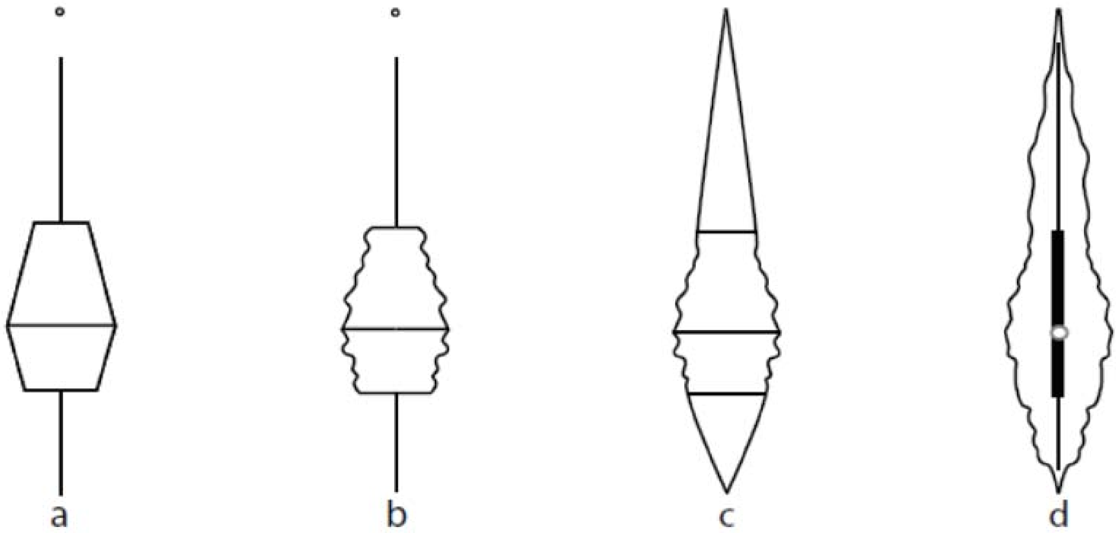
Box-and-whisker plots

- IQR = 44-29 = 15
- Upper Fence = 44 + 15*1.5 = 66.5
- Lower Fence = 29 - 15*1.5 = 6.5



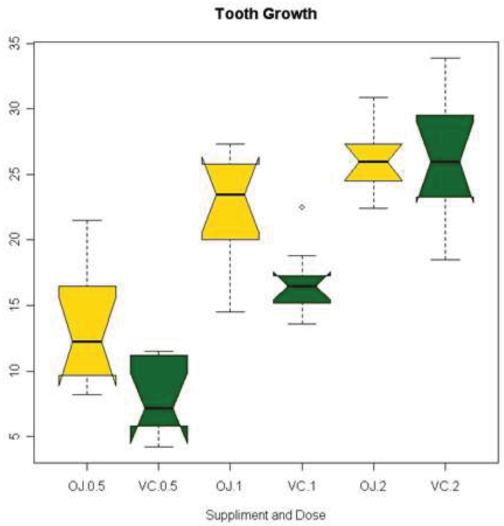
Extensions to box-plots

Compared to the classical box plot, what extra information is provided in the plots below?



Extensions to box-plots

- Several methods exist for adding density to the box plot:
  - a) histplot,
  - b) vaseplot,
  - c) box-percentile plot,
  - d) violin plot (a combination of a boxplot and a kernel density plot)



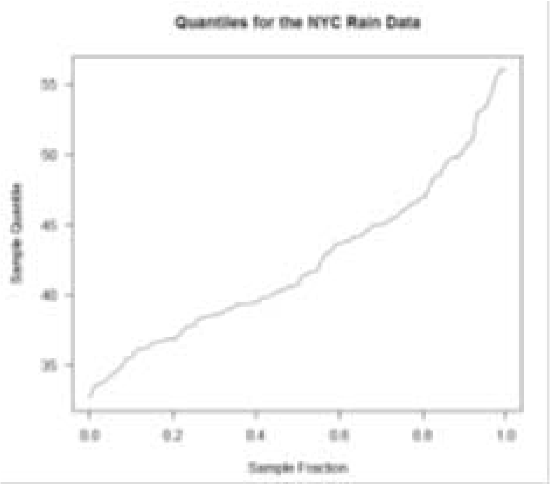
- In the notched boxplot, if the notches of two boxes do not overlap this is “strong evidence” for their medians to be different

Quantile-Quantile (Q-Q) Plots

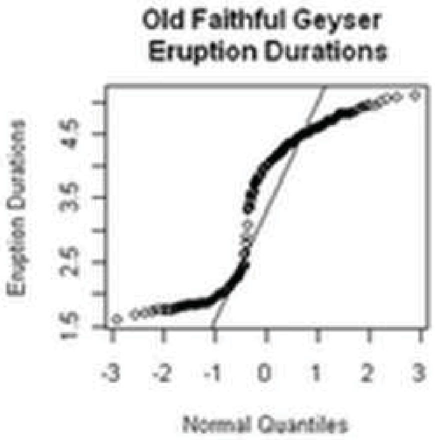
- We have seen a quantile function before . . . ; one corresponding to a normal density function
- In general, the **quantile function** of a probability distribution f is the inverse of its cumulative distribution function (cdf) F
- Quantile functions as well can be estimated from the data at hand.
- If we consider the estimated quantile function to be a "good" estimate (sample level) for the truth (population level), it will learn us something about the true underlying mechanisms of the data
- If we assume a particular "model" or mechanism that could have generated the data, we can compare the quantile function corresponding to this "theoretically proposed" distribution to the quantile function corresponding to our observed data

Q-Q Plots

Quantile plots: The sample quantiles are plotted against the fraction of the sample they correspond to



Q-Q plots: The sample quantiles are plotted against the theoretical quantiles ("observed" quantiles are compared to "expected" quantiles under the assumed model)



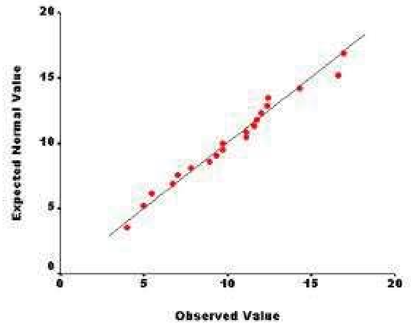
Q-Q Plots

- In general, Q-Q plots allow us to compare the quantiles of two sets of numbers. They go beyond the information provided by box-plots (also using quantiles), in that Q-Q plots give us a clue about the validity of a proposed model for the data or data generation mechanism
- There is a cost associated with this extra detail. We need more observations than for simple comparisons
- Remark:
  - A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function.
  - Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions

Examples (I)

Normal Q-Q Plots of Samples from Normal Populations

- A sufficiently trained statistician can read a Q-Q-plot like a holistic medical doctor can read the internal organs of a person. Interpreting Q-Q plots is more a visceral than an intellectual exercise. The uninitiated are often mystified by the process.
- Experience is the key here. The first step is to examine **normal Q-Q plots** of samples known to be from normal populations, to get some idea of how much straggling about the line is acceptable.



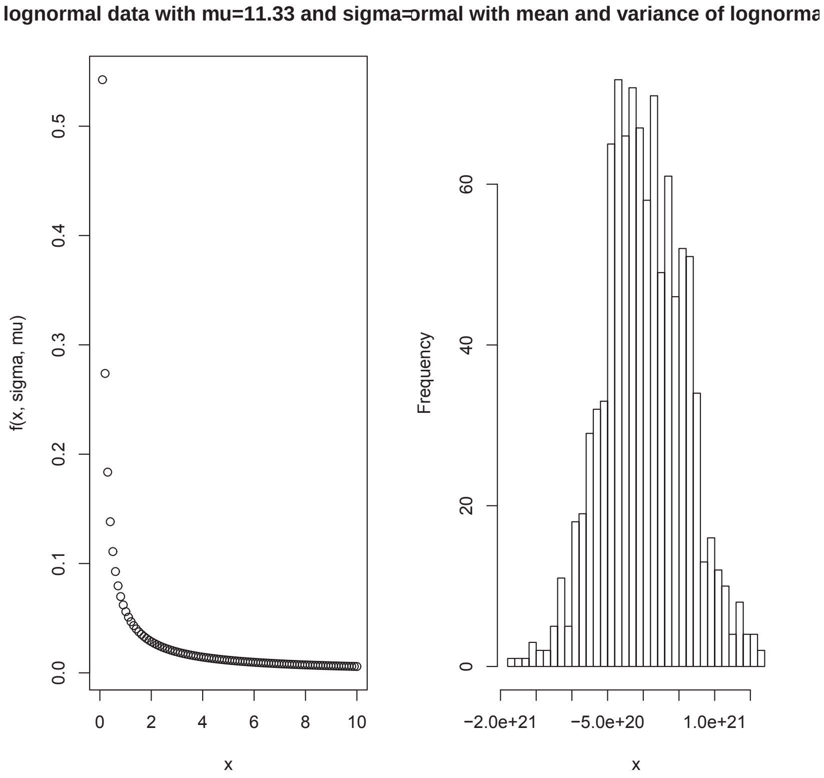
Normal Q-Q plot of a sample of 20 observations from a Normal population with mean 10 and standard deviation 3

Examples (II)

Normal Q-Q Plots of Samples from Skew Populations

- The lognormal density is given by

$$f(x) = 1/(\sqrt(2\pi)\sigma x)e^{-((\log x - \mu)^2/(2\sigma^2))}$$



Examples (II)

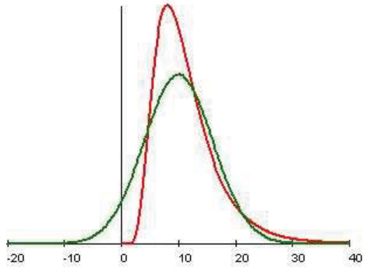
Normal Q-Q Plots of Samples from Skew Populations

- Specific departures from normality in the population being sampled manifest themselves as specific departures from the straight and narrow in the Q-Q plot.
- If the population being sampled is actually skewed to the right, i.e. has a long right hand tail, and thus short left tail, then the sample quantiles close to 1 will lie to the right of where normality would place them, and similarly for the sample quantiles close to 0. For quantiles closer to 0.5, the Normal quantiles will exceed those of the sample.

Why?

Why Do Q-Q Plots from Skewed Populations Look Like They Do?

- A plot of a right-skewed lognormal population compared to a normal population with the same mean and standard deviation (mean=11.33, st.dev.=6.04):

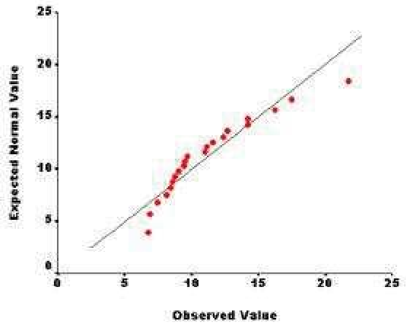


- If you imagine where the 0.1, 0.2,...,0.9 quantiles (for example) lie in both populations, it seems reasonably clear that the normal quantiles will be less than their lognormal counterparts to begin with. By quantile 0.5 the situation has been reversed, since the lognormal has a median of 10 and the Normal a median of 11.33. At the right hand end of the plot, the normal quantiles will again be less than their lognormal counterparts.

Examples (II)

Normal Q-Q Plots of Samples from Skew Populations

- The result is a Q-Q plot which resembles the left hand top of an arch, starting below the target line (or to the right if you prefer), arching across it and then back to finish below (or to the right of) the line again.
- If the sampled population is skewed to the left, the arch is reflected about $Y = X$, starting above it, crossing below and then back to finish above.



Normal Q-Q plot of a sample of 20 observations from a lognormal population with mean 10 and standard deviation 3. This population is skewed to the right (i.e. it has a long right hand tail).

Examples (III)

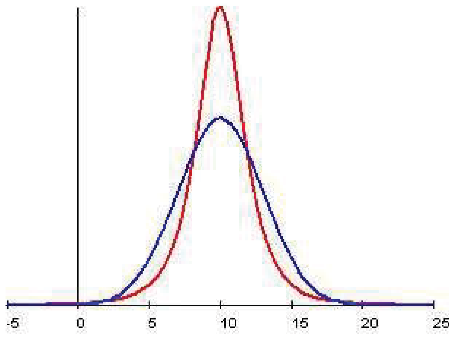
Normal Q-Q Plots of Samples from Heavy Tailed (Leptokurtic) Populations

- Heavy tailed populations are symmetric, with more members at greater remove from the population mean than in a Normal population with the same standard deviation.
- To compensate for the extreme members of the population, there must also be higher concentration around the population mean than in a Normal population with the same standard deviation. Heavy tailed populations have higher, narrower peaks than the benchmark Normal population. Hence, the term leptokurtic - narrow arched.

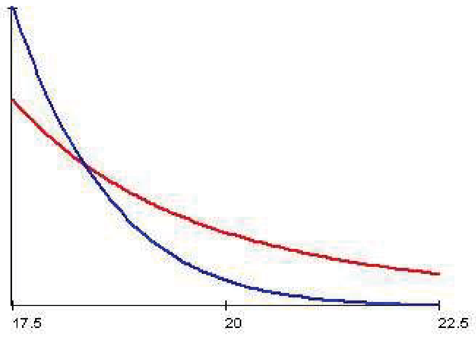
Examples (III)

Normal Q-Q Plots of Samples from Heavy Tailed (Leptokurtic) Populations

- With a normal population (same mean and standard deviation as the leptokurtic population) as benchmark, the sample quantiles might be expected to start ahead of their normal counterparts, but be soon overtaken by them. Symmetry would place both sample and target median back together again. The situation would be reversed as you move from the median into the right hand tail, with the sample quantiles in front of the targets to begin with, but eventually being overtaken by them.



Heavy tailed population (red) compared with a normal population (blue)

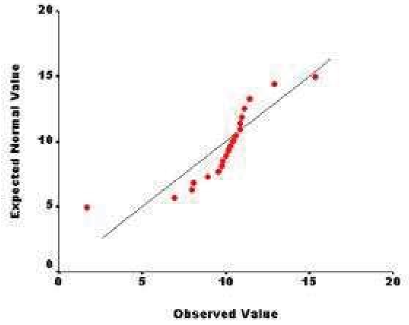


A section of the extreme right hand tails of the two populations, showing the extended reach of the heavy tailed (red) population compared to its Normal (blue) benchmark.

Examples (III)

Normal Q-Q Plots of Samples from Heavy Tailed (Leptokurtic) Populations

- The result is a Q-Q plot which resembles a stretched S, starting to the left of the target line, and ending to the right of it, having crossed it three times in between.



Normal Q-Q plot of a sample of 20 observations from a heavy tailed population with mean 10 and standard deviation 3.

Methods of EDA: two-way

- 2 Categorical Variables
  - Frequency table
- 1 Categorical, 1 Continuous Variable
  - Stratified stem-and-leaf plots
  - Side-by-side box plots
- 2 Continuous variables
  - Scatterplot

2 categorical variables

Frequency Table

Age Interval	Gender		Total
	Female	Male	
20-29	1	2	3
30-39	2	2	4
40-49	1	1	2
50-59	1	0	1
Total	5	5	10

It looks like the men tend to be younger than women in this example.

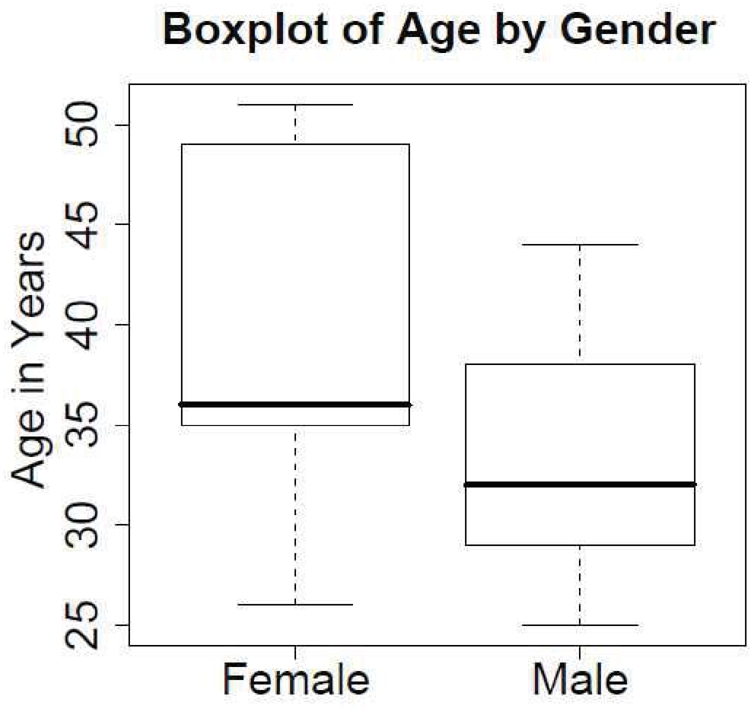
1 categorical + 1 continuous variable

Stratified Stem-and-Leaf plots

Female		Male		
Age Interval	Obs.	Age Interval	Obs.	
20-29	6	20-29	5 9	
30-39	5 6	30-39	2 8	
40-49	9	40-49	4	
50-59	1	50-59		
Total	5		5	10

1 categorical + 1 continuous variable

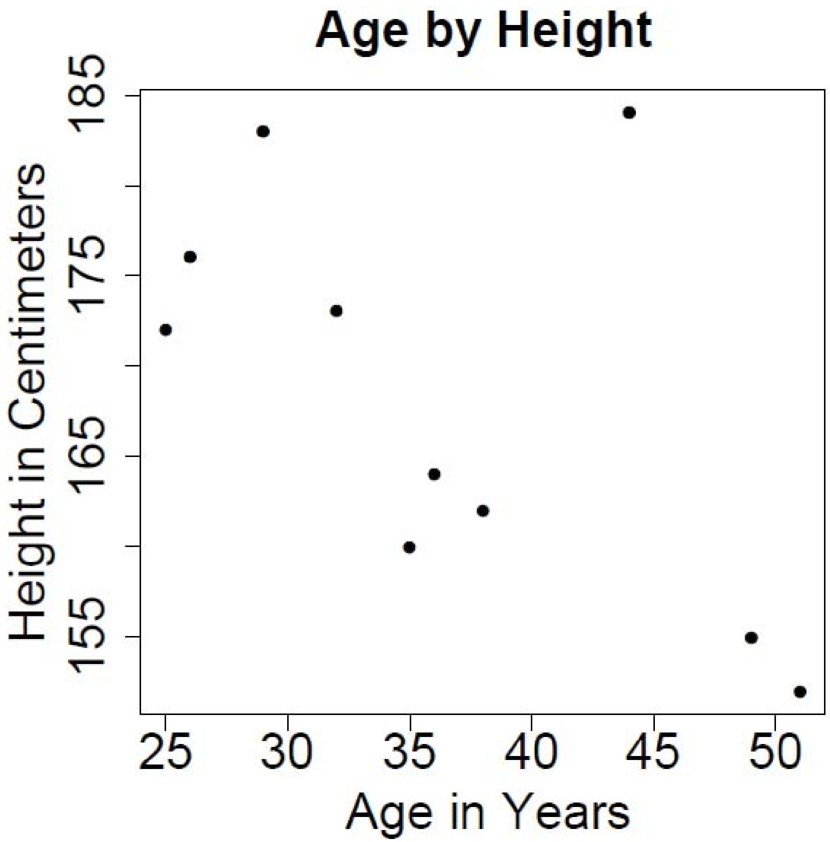
Side-by-Side Box Plots



Allows us to compare the distribution of the continuous variable (age) across values of the categorical variable (gender)

2 continous variables

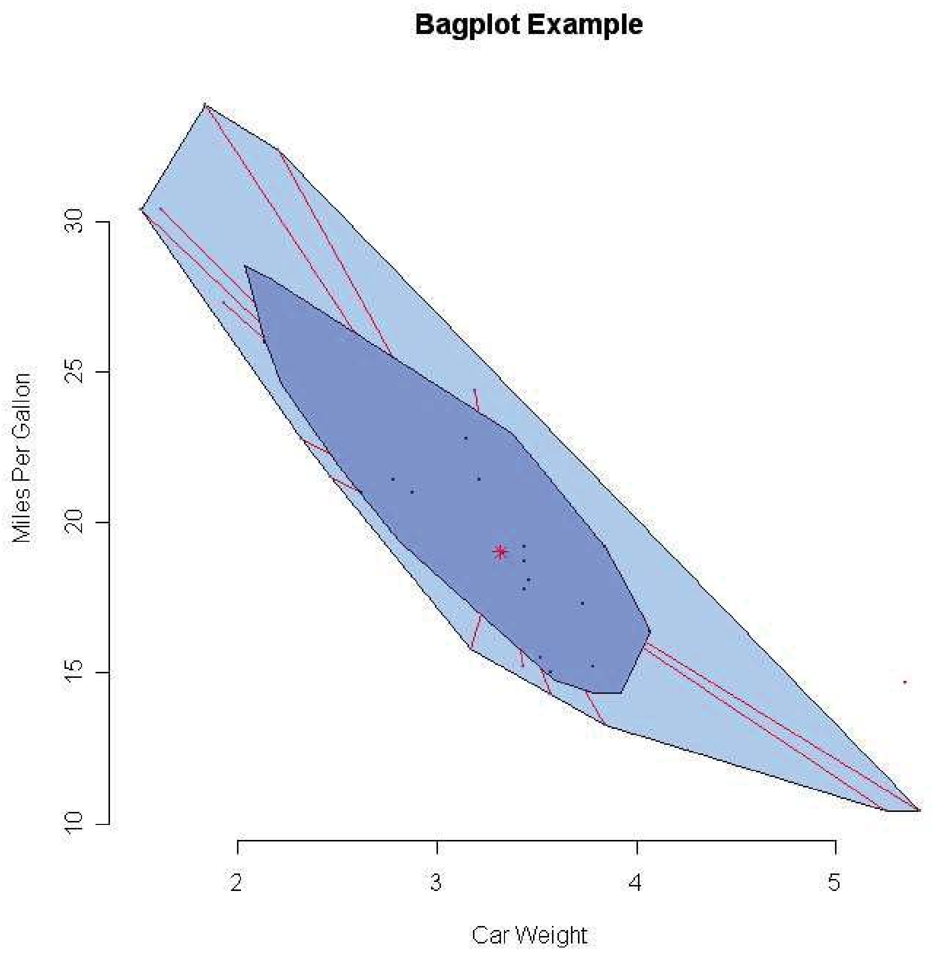
Scatterplot



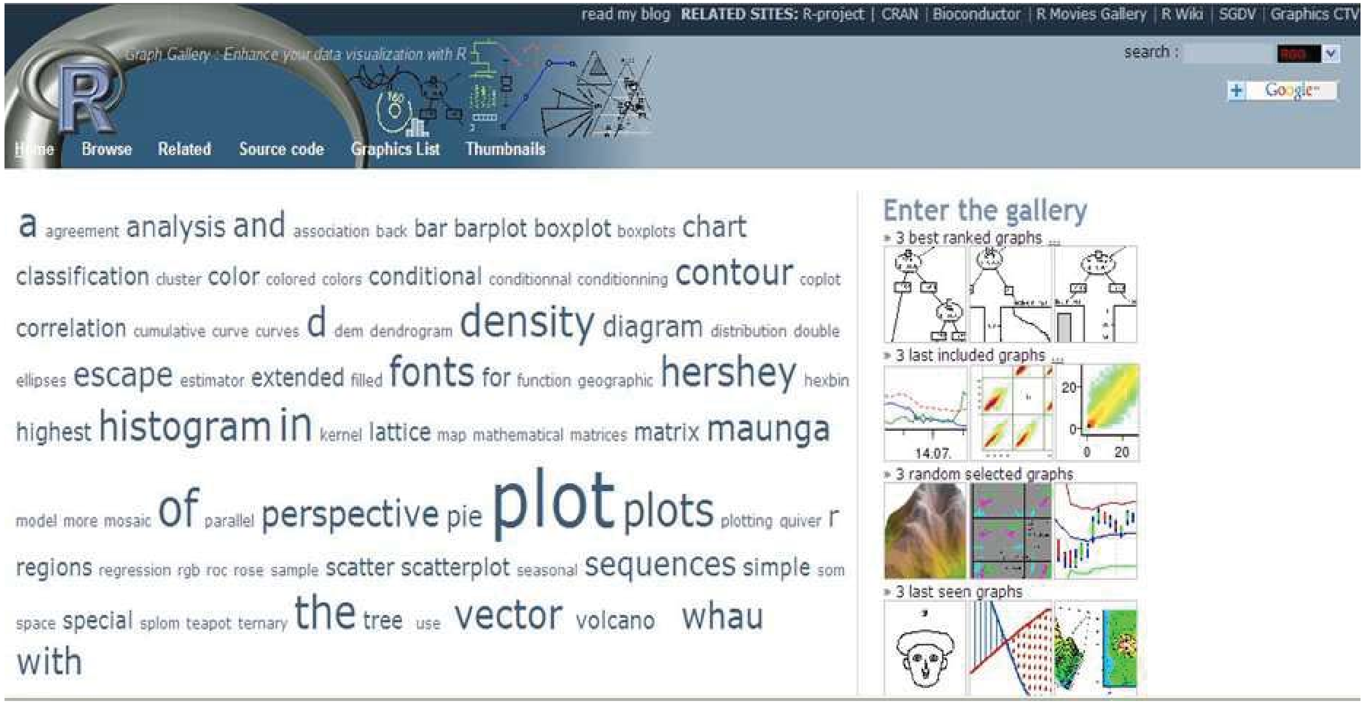
Scatterplots visually display the relationship between two continuous variables

2 continous variables

```
library(aplpack) attach(mtcars) bagplot(wt,mpg, xlab="Car Weight
ylab="Miles Per Gallon", main="Bagplot Example")
```



R gallery


read my blog RELATED SITES: R-project | CRAN | Bioconductor | R Movies Gallery | R Wiki | SGDV | Graphics CTV
Graph Gallery : Enhance your data visualization with R
search :
Google
Home
Browse
Related
Source code
Graphics List
Thumbnails

a agreement analysis and association back bar barplot boxplot boxplots chart classification cluster color colored colors conditional conditionnal conditionning contour coplot correlation cumulative curve curves d dem dendrogram density diagram distribution double ellipses escape estimator extended filled fonts for function geographic hershey hexbin highest histogram in kernel lattice map mathematical matrices matrix maunga model more mosaic of parallel perspective pie plot plots plotting quiver r regions regression rgb roc rose sample scatter scatterplot seasonal sequences simple som space special splom teapot ternary the tree use vector volcano whau with

Enter the gallery

» 3 best ranked graphs ...

» 3 last included graphs ...

» 3 random selected graphs

» 3 last seen graphs

Assumptions of EDA

- Virtually any data analysis approach relies on assumptions that need to be verified
- There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand ”behave like”:
  - random drawings,
  - from a fixed distribution,
  - with the distribution having fixed location and
  - with the distribution having fixed variation
- The data are called to follow a **univariate process**

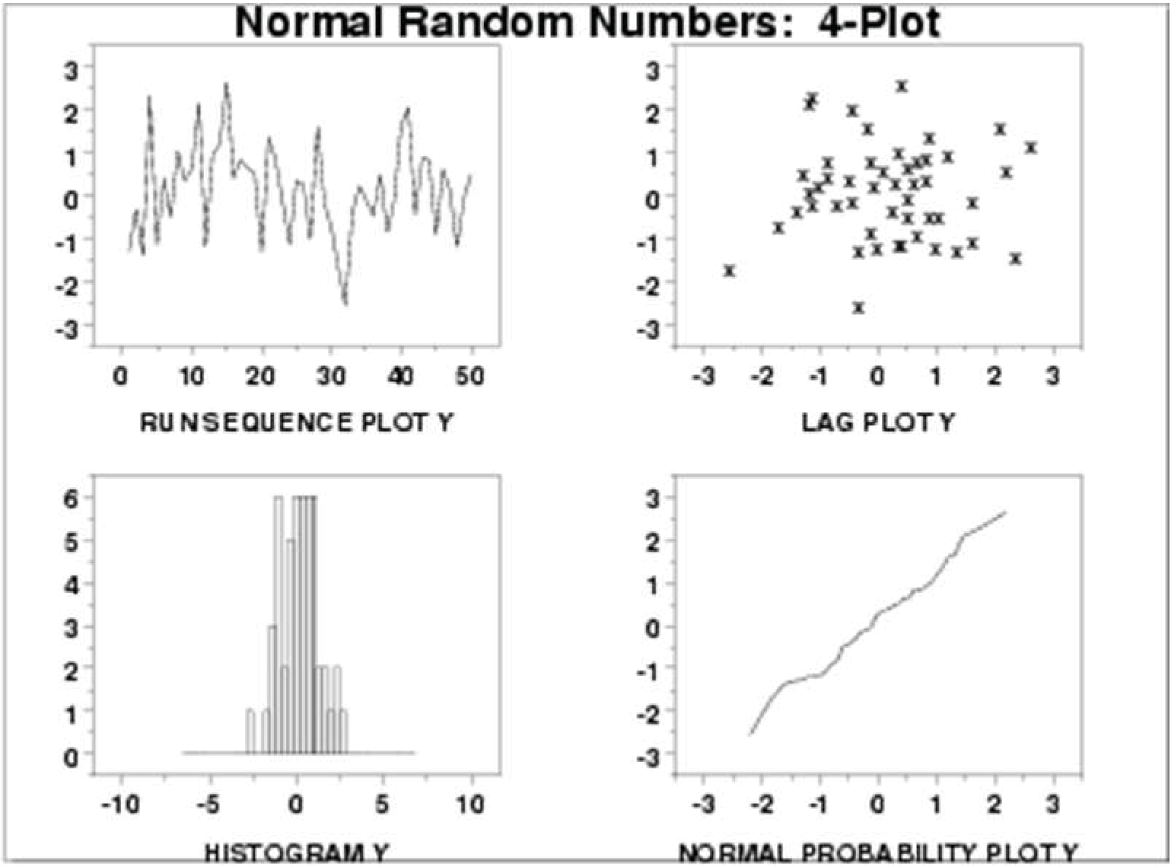
Assumptions of EDA

- The most common assumption in any data analysis is that the differences between the raw response data and the predicted values from a fitted model (these are called **residuals**) should themselves behave like a univariate process
- So, if the residuals from the fitted model behave like the ideal, then testing the underlying assumptions for univariate processes becomes a tool for the validation and quality of fit of the chosen model.
- On the other hand, if the residuals from the chosen fitted model violate one or more of the aforementioned univariate assumptions, then we can say that the chosen fitted model is inadequate and an opportunity exists for arriving at an improved model.

Verifying the Assumptions of EDA

- The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:
  - **run sequence plot** (Y_i versus i) – upper left on next slide
  - **lag plot** (Y_i versus Y_{i-1}) – upper right on next slide
  - **histogram** (counts versus subgroups of Y) – lower left on next slide
  - **normal probability plot** (ordered Y versus theoretical ordered Y) - lower right on next slide
- Together they form what is often called a 4-plot of the data.

Verifying the Assumptions of EDA



Interpretation of 4-plots

- **Randomness**: If the randomness assumption holds, then the lag plot $(Y_i$ versus $Y_{i-1})$ will be without any apparent structure and random.
- **Fixed Distribution**: If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then the histogram will be bell-shaped, and the normal probability plot will be linear.
- **Fixed Location**: If the fixed location assumption holds, then the run sequence plot $(Y_i$ versus $i)$ will be flat and non-drifting.
- **Fixed Variation**: If the fixed variation assumption holds, then the vertical spread in the run sequence plot $(Y_i$ versus $i)$ will be the approximately the same over the entire horizontal axis.

Can we reverse the reasoning?

Interpretation of 4-plots

- **Run Sequence Plot**: If the run sequence plot (Y_i versus i) is flat and non-drifting, the fixed-location assumption holds. If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.
- **Lag Plot**: If the lag plot is without structure, then the randomness assumption holds.
- **Histogram**: If the histogram is bell-shaped, the underlying distribution is symmetric and *perhaps* approximately normal.
- **Normal Probability Plot**: If the normal probability plot is linear, the underlying distribution is approximately normal

When assumptions are violated

- Consequences of non-randomness:
  - All of the usual statistical tests are invalid.
  - The calculated uncertainties for commonly used statistics become meaningless.
  - The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
  - Even the simple model linear regression model becomes invalid.
  - The parameter estimates become suspect and non-supportable
  - ...
- When violations cannot be corrected in some sense, usually a more complicated analysis strategy needs to be adopted (for instance: mixed modelling to account for dependencies caused by multiple measurements taken over a specific time span, for the same individual).