Introduction

Sampling Objectives:

- Sampling ultimately aims at gaining information about of the ***underlying population*** by using a sample. The information could be used to:
    - approximate the distribution of the whole population
    - or to obtain some of its distributional parameters (e.g. central tendency, variability etc.)
- General questions:
    - How do we obtain a ***representative sample*** from the population?
    - ***Which*** objects and ***how many*** objects must be included into the sample to provide ***on average an accurate snapshot*** of the total underlying population.
- ***Lack representativeness*** leads to a ***biased sample***.
  $\Rightarrow$ Example: Industrious students are more likely to be on campus. Random interviews on campus have the tendency to oversampling this group.
- The information obtained from ***randomly*** sampled observations always will ***deviate to some degree*** from the underlying population
  $\Rightarrow$ we just can ***aim*** at making this deviation ***small and well balanced***.
- Impact of ***sample size***: The likelihood of obtaining a representative sample increases as we ***sample a larger cross-section*** of the population.
  Cross-section means that no particular sub-group of the population should be favored to be

included into the sample.
Counterexample: Estimation of election outcomes in the "big data" article. The *Literary Digest* only sampled affluent people about their presidential election attitudes.

- ***Uncertainty*** is the price we pay for not fully enumerating every object of the underlying population
- Def. Sampling Error: Sampling error is the ***uncertainty*** that arises by working with a (random) sample rather than with the entire population
- Def. Sampling Bias: Sampling bias occurs when the procedure used to draw sample observations ***selectively favors the inclusion or suppression*** of specific population members. If the sampled members systematically differ from the overall population then the sample cannot be representative.
- Sampling bias can be avoided by implementing an appropriate ***sampling plan*** and ***check for recording errors*** of the sampled data.

The relevant population under investigation

- Key question: ***for which exact population*** do we want to make ***inferential statements***? This is a decision of inclusion and exclusion of objects into the ***sampling frame***.
  - A strict and operationally useful definition of the population for which we want to make a statement is required.
  - Each member in the population must have an ***equal chance*** of entering the sample.

- Consider also the costs, time and the geographical limitations to sampling.

Sampling Design

- Def. Sample Design: A sample design describes the specific ***procedure used to select objects*** from the sampling frame.
- In inferential statistics, inclusion of objects into a sample is done by some ***random*** procedure. Randomness will ***control for any biases by balancing*** the over- and under-representativeness of specific sub-groups.
- The ***probability of inclusion*** into the sample for each object in the sampling frame must be known before the sample is drawn.
  If particular objects, which happen to share common properties, have a higher inclusion probability then these objects may induce a biased sample.

Simple Random Sampling

- Def. Probability Sample: The probability of any individual member of the population being picked into the sample can be determined
- Def. Simple Random Sample and Finite Population: A simple random sample from a finite population of size N is one in which each possible sample object has an ***equal selection probability***.

- Simple random sampling does not rule out the ***chance of obtaining extreme sample observations***. However, since we know the selection probability we can calculate the ***likelihood of obtaining an extreme a sample***.
- With increasing sample size the probability of obtaining an extreme sample is decreasing.

- Example: Enumeration of all combinations of 2 objects out of a population of 4, i.e., $\binom{4}{2} = 6$

- Problem: Sampling without replacement leads to ***statistically dependent draws***.
  $\Rightarrow$ The probability for the selecting the second observation changes after the first observation has been selected.

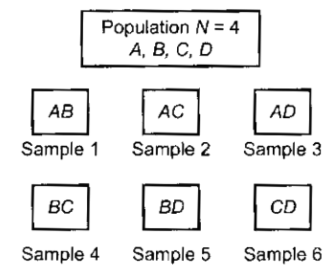


FIGURE 6-3. Samples of size $n = 2$.

- Rather than enumerating all possible combinations as in Figure 6-3,
  - we can assume for ***large enough*** populations that each member in the population has an equal probability of being selected into the sample
  - Therefore, the individual draws of sample members become approximately statistically independent among each other.

    Thus, the probability of any ***set of sampled members*** can be approximated by

$$\Pr(\omega_1 \cap \omega_2 \cap \cdots \cap \omega_n) = \Pr(\omega_1) \cdot \Pr(\omega_2) \cdots \Pr(\omega_n) = \pi^n,$$

which assumes independence of the each object being drawn into the sample.

Sampling Distributions of Statistics

- Recall: the population parameters are denote by Greek characters, e.g., μ and σ^2, and the sample statistics are denoted by Latin letters or have a hat on top, e.g., means $\overline{x}$ and s^2 or $\hat{\mu}$ and $\hat{\sigma}^2$, respectively.
- <u>Def. Sample Statistic:</u> A sample statistic is itself a ***random variable*** – based on the random variables $X_1, X_2, \ldots, X_n$ in the sample – that ties these individual random variables together through some functional expression.
  - Example: the sample statistic function is $\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ and its random outcome for a particular sample becomes $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$.

Sampling Distributions

- <u>Def. Sampling Distribution of a Statistic:</u> A sampling distribution is a probability distribution of a sample statistic.

  That is, the sample statistic must have a distribution because it is calculated from a set of random variables.

- The sampling distribution of a statistic can be, in theory, developed by taking ***all possible*** samples of size *n* from a population, calculating the values of the sample statistic for each of these sampling outcomes and drawing the distribution of these values.
- Example: Evaluate the ***distribution of the sample mean*** based on $n = 3$ random draws without replacement and irrespectively of the order from a population of size $N = 5$.

Element	Values of x
A	$x_1 = 6$
B	$x_2 = 6$
C	$x_3 = 5$
D	$x_4 = 4$
E	$x_5 = 4$

TABLE 7-2
Possible Samples of Size $n = 3$ from Population $N = 5$

Elements in sample	Values of X	Mean $\bar{x}$
ABC	6, 6, 5	5.7
ABD	6, 6, 4	5.3
ABE	6, 6, 4	5.3
ACD	6, 5, 4	5.0
ACE	6, 5, 4	5.0
ADE	6, 4, 4	4.7
BCD	6, 5, 4	5.0
ACE	6, 5. 4	5.0
BDE	6, 4, 4	4.7
CDE	5, 4, 4	4.3

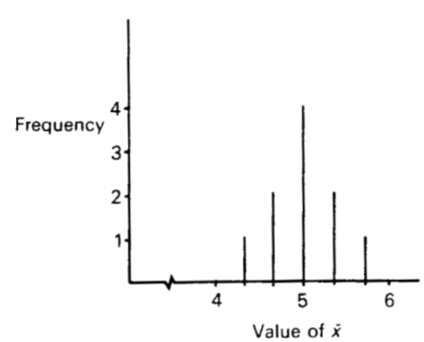


FIGURE 7-6. Sampling distribution of $\bar{X}$.

- This distribution can be characterized by its expected value

$$E(\bar{X}|n = 3, N = 5) = \frac{1}{10} \cdot 4.3 + \frac{2}{10} \cdot 4.7 + \frac{4}{10} \cdot 5.0 + \frac{2}{10} \cdot 5.3 + \frac{1}{10} \cdot 5.7 = 5.0$$

and analogue for its variance.

- If the expected value of the sampling sample statistic is ***equal*** to the expectation of the population then the sampling statistics is said to be ***unbiased***.

  We usually prefer statistical estimation rules that are unbiased.

- The standard deviation of the sampling statistics is called the ***standard error***. For the mean statistic its standard error is denoted by $s_{\bar{X}}$.
  The standard error measures the uncertainty that the sample statistic will deviate from its expected value.
  We prefer statistical estimation rules that lead to small standard errors (i.e., small uncertainty).
- General example:

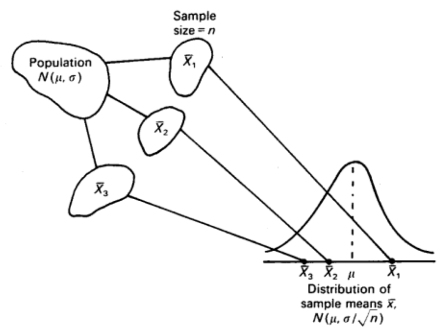


FIGURE 7-7. Central limit theorem and the distribution of sample means.

Central Limit Theorem

- Def. Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be a ***random independent*** sample of size *n* drawn from an *arbitrarily* distributed population with expectation μ and standard deviation σ. Then for large enough sample sizes *n*, the sampling distribution of $\bar{X}$ is asymptotically (i.e., as $n \to \infty$) normal distributed with $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
- There are two part to this theorem:

1. Irrespectively of the sample size the expected value of $\bar{X}$ is $E(\bar{X}) = \mu$ and the variance is $Var(\bar{X}) = \sigma^2/n$.
   Note, n in the denominator. Therefore, as the sample size n increases the standard error $s_{\bar{X}} = \sqrt{\sigma^2/n} = \frac{\sigma}{\sqrt{n}}$ of the mean will shrink.
2. Asymptotically the sample mean will follow a normal distribution irrespectively of the underlying distribution of the population.

Proof for independent sample objects: $Var\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \underbrace{Var(X_i)}_{=\sigma^2} = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$

- <u>Example:</u> Variability of Sample Statistics:
- <u>Example:</u> Central limit theorem with the -script

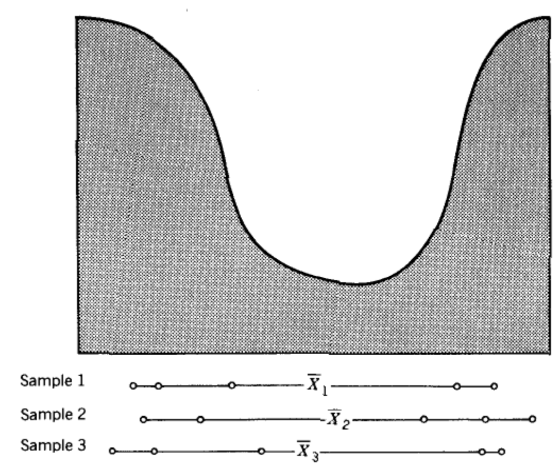
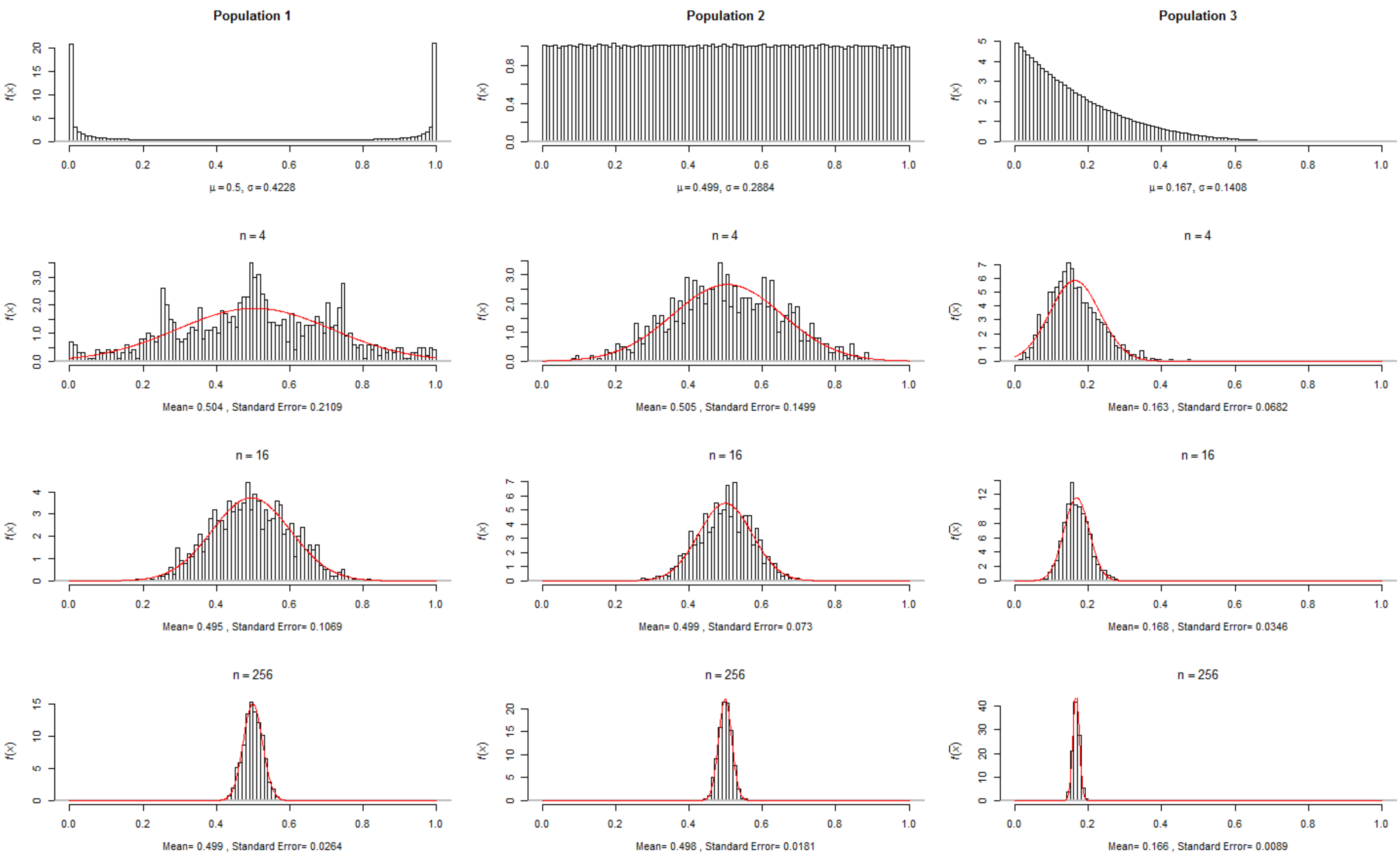


FIGURE 2.24 Three samples of five observations drawn randomly from U-shaped distribution. Means of samples are shown by $\bar{X}$.

CENTRALLIMIT.R:



- Different underlying population distributions will lead to a normal distribution of the sample statistic $\bar{X}$ (and the sum $\sum_{i=1}^{n} X_i$), which is unbiased and the standard error shrinks with increasing sample size following the rule $\sigma_{\bar{X}} = \frac{\sigma_{population}}{\sqrt{n}}$.

Random Sampling in Using Computer Algorithms

- Random sampling procedure plays an important role in modern statistics and machine learning. See, for instance, BBR pp 418-426 on Bootstrapping.
- The script **`BasicRandomNumberGenerator.R`** implements the simple pseudo random number algorithm described in BBR on pp 364-265.
- See the document **`RNGVersions.pdf`** for a further discussion.

Random Sampling Designs

- The objective of sampling theory is to develop sampling plans and statistics that lead to the ***most precise estimators of population properties***, i.e., estimators with low uncertainty (standard error) and control for any biases.
- This is an ***optimization problem***, perhaps, under ***constraints***.
- ***Weighting*** can control for biases:
  - the impact of observations with a higher probability of being selected into the sample need to be ***weighted down*** and

- the impact of observations with lower selection probability needs to be ***weighted up***

This may achieve representativeness.

Stratified Random Sampling (statistical details not test relevant)

- Def. Stratified Random Sampling: A stratified random sample is obtained by
  [1] splitting the population into k preferably ***homogeneous*** classes – ***also called strata*** – and
  [2] selecting a simple random samples of a predetermined size n_j from each stratum j.
- The sample statistics from these strata-specific samples are then combined into the global sample statistic. Each stratum will have a ***stratum-specific weight***.
- Homogeneity assumption:
  - External knowledge is required to split the population into k preferably ***homogenous*** strata.
  - Homogeneity relates the ***measured attributes*** of the sample observations within each stratum.
  - External proxy variables that are **closely correlated** with the measure attributes can be surrogates.
  - Homogeneity means that the ***variances of the sub-populations*** within each stratum are ***less*** than the overall population variance.
- Advantage: The additional control leads to a reduction in the overall sampling error
- The underlying strata-specific data structure becomes:

Strata	1	2	$\cdots$	k
Size	N_1	N_2	$\cdots$	N_k
Variance	σ_1^2	σ_2^2	$\cdots$	σ_k^2
Sampling cost per unit	c_1	c_2	$\cdots$	c_k
Population	$\left\{X_{1,1}, X_{2,1}, \ldots X_{N_1,1}\right\}$	$\left\{X_{1,2}, X_{2,2}, \ldots X_{N_2,2}\right\}$	$\cdots$	$\left\{X_{1,k}, X_{2,k}, \ldots X_{N_k,k}\right\}$
Variable sample size	n_1	n_2	$\cdots$	n_k

- Challenges of Stratification:
  - We must have some ***external knowledge*** about the population characteristics to stratify it properly so that the internal strata variances become minimal
  - The strata membership for each object in the population must be known.
  - The sub-population size in each stratum must be known.
  - We should have a rough idea of the costs of obtaining a sample observation from each stratum. These costs may vary from strata to strata.
- The estimates of
  - the ***strata means*** are $\bar{x}_j = \frac{1}{n_j} \cdot \sum_{i=1}^{n_j} x_{i,j}$ and
  - the ***strata variance*** are $s_j^2 = \frac{1}{n_j - 1} \cdot \sum_{i=1}^{n_j} \left(x_{i,j} - \bar{x}_j \right)^2$
- The ***overall*** estimated mean and variance become ***weighted estimates*** of the strata statistics
  - $\bar{x}_{overall} = \frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot \bar{x}_j$ and
  - $s_{overall}^2 = \underbrace{\frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot s_j^2}_{\text{within strata variation}} + \underbrace{\frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot \left(\bar{x}_j - \bar{x}_{overall} \right)^2}_{\text{between strata variation}}$, respectively.
  - ***Not weighting*** leads to ***biased*** overall estimates.
- The total population size is $N = N_1 + N_2 + \cdots + N_k$ and $N_j \geq 2$.
- The stratum-specific sample size n_j needs to satisfy the constraints:

- $0 \le n_j \le N_j \quad \forall\ j \in \{1,2,\ldots,k\}$
- n_j needs to be integer numbers.
- for large N_j usually sampling ***with replacement*** is assumed to make calculations easier.

* The optimization problem to determine the best strata sample sizes $\{n_1^*, n_2^*, \ldots, n_k^*\}$
  - The optimal sampling plan $\{n_1^*, n_2^*, \ldots, n_k^*\}$ for $\bar{x}_{overall} = \frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot \bar{x}_j$ can be analytically determined by ***minimizing the standard error of the global estimator (objective function)***

$$\min_{n_1, n_2, \ldots, n_k} Var(\underbrace{\tfrac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot \bar{x}_j}_{\bar{x}_{global}}) = \sum_{j=1}^{k} \left(N_j / N\right)^2 \cdot \frac{\sigma_j^2}{n_j}$$

subject to the cost constraint

$$c_{total} \equiv c_0 + \sum_{j=1}^{k} c_j \cdot n_j^* .$$

  - The solution to this optimization problem under constraints can be obtained with the *Lagrange Multiplier* technique. It becomes

$$n_j^* = \left(c_{total} - c_0\right) \cdot \frac{N_j \cdot \sigma_j \big/ \sqrt{c_j}}{\sum_{j=1}^{k} N_j \cdot \sigma_j \cdot \sqrt{c_j}}$$

However, the n_j^* must be rounded to the closest integer value $n_j^* \geq 2$.

Another optimization technique called *Integer Programming* would give exact results, but it does not provide an analytical solutions.

- General rules for sample size selection n_j^* in each stratum: Select a larger sample size n_j^* in strata j
  - $n_j^* \uparrow$ if $N_j \uparrow$: if strata j consists of a larger proportion N_j/N of the population
  - $n_j^* \uparrow$ if $\sigma_j \uparrow$: if strata j is more heterogeneous (larger internal variance σ_j^2)
  - $n_j^* \uparrow$ if $c_j \downarrow$ if it is less expensive to sample in stratum j (small c_j)
- Stratification can also be used to oversample otherwise underrepresented groups.

Cluster Random Sampling (statistical details not test relevant)

- Def. Cluster Random Sampling: In clustered random sampling the population is divided by convenience into mutually exclusive classes. In a two-steps procedure:
  - randomly a ***subset of clusters*** are picked.
  - a specific number of ***observations are sampled*** from the selected clusters.
- Clusters are supposed to be ***heterogeneous*** (strong mixing of attribute values) with regards to the attributes under investigation (opposite to stratified sampling).
  This means we expect that each ***cluster is being representative*** of the whole population.
- The overall mean and variance estimates in clustered sampling again become

- $\bar{x}_{overall} = \frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot \bar{x}_j$ and
- $s_{overall}^2 = \frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot s_j^2 + \frac{1}{N} \cdot \sum_{j=1}^{k} N_j \cdot \left(\bar{x}_j - \bar{x}_{overall}\right)^2$, respectively.

- Therefore, we need to know at least the cluster sizes N_j.
- Clustered sampling can save sampling costs but bears the potential of high sampling error and sampling bias.

Discussion

- With regards to the ***sampling error*** stratified sampling is the most efficient and clustered sampling the least efficient.
- The sampling error of random and systematic random sampling lies in-between stratified and clustered sampling.
- Requirements for *a priori* knowledge of the sampling frame are the highest for stratified sampling.

Spatial Sampling

Spatial sampling issues are explicitly discussed in the GISC course *Pattern Analysis*.

Spatial Point Sampling Approaches:

- For a countable number of given spatial objects (such as a finite set of points) standard sampling procedure can be applied.

- For spatially continuous surfaces ***random locations*** need to be generated:
  - How to pick a random point from a square study area $\mathbb{R}$?
    Select the *x*-coordinate from a uniform distribution $X_i \sim \mathcal{U}(x_{min}, x_{max})$ and the *y*-coordinate from $Y_i \sim \mathcal{U}(y_{min}, y_{max})$, respectively.
  - The resulting ***local densities*** of the sample points are approximately ***uniform*** (i.e., constant).
  - This sampling procedure leads to ***complete spatial randomness***.
  - Example: Use complete spatial randomness to estimate areas (areal integrals) with the script **`EstimateAreaBySampling.r`**.
- For systematic, stratified, or clustered sampling the reference frame can either be a square raster cells or a hexagonal grid.

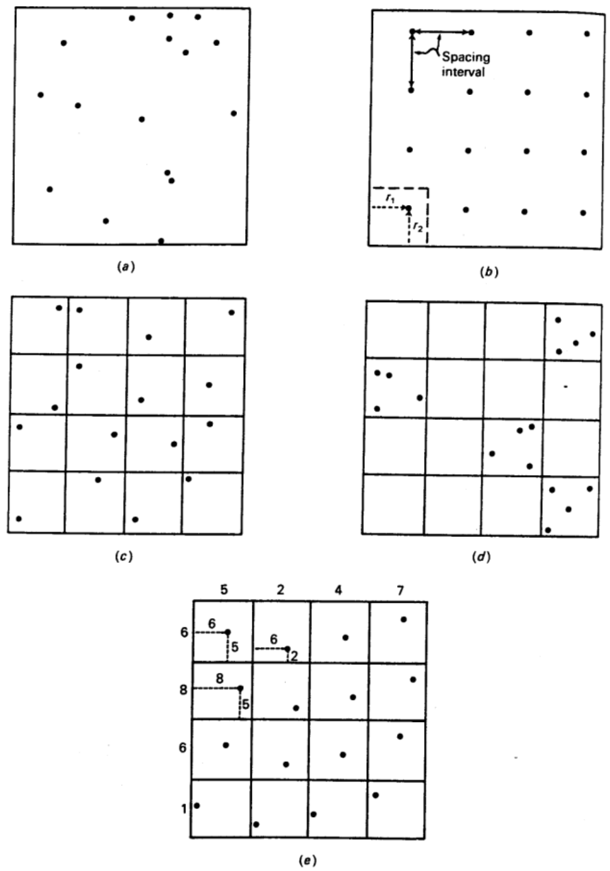


FIGURE 7-17. (a) A simple random point sample; (b) a systematic areal sample; (c) an areally stratified random sample; (d) a cluster sample; (e) a stratified, systematic, unaligned areal sample.

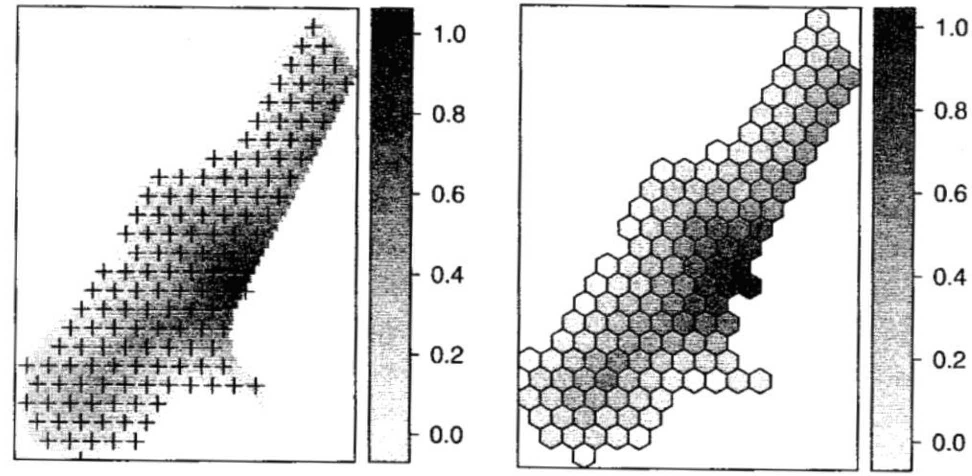


Fig. 6.2. Hexagonal points (*left*) and polygons (*right*)

- Hexagons have the advantage that the ***nearest neighbor points*** are all ***equidistant***.
- This is not the case for grid cells were ***diagonal cell*** centers are further apart than ***horizontal and vertical cell*** centers.

Quadrate Sampling

- Quadrates of a given size are randomly distributed over the map

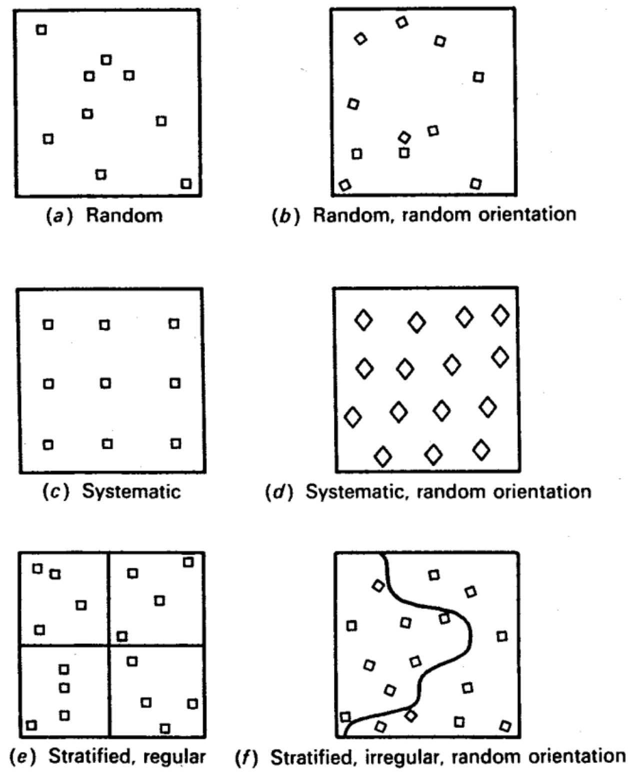


FIGURE 7-14. Sampling designs for quadrat sampling.

Traverse Sampling

- We want to sample along line segments with a total length of *L* (replaces the sample size *n*).
  - These segments are defined by random starting and ending points along the study area's boundary.

- Then randomly sample a sub-length l_i along the traverse
- Repeat process until $L = \sum_{i=1}^{n} l_i$.

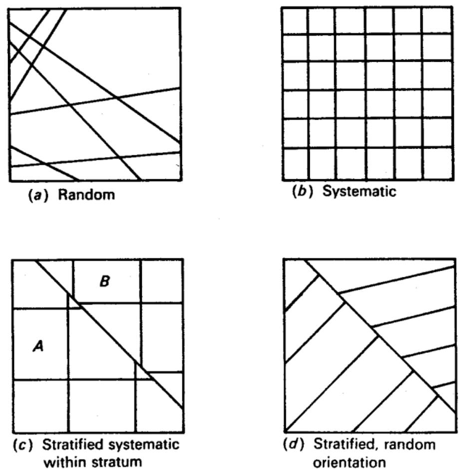


FIGURE 7-15. Sampling designs for traverse sampling.

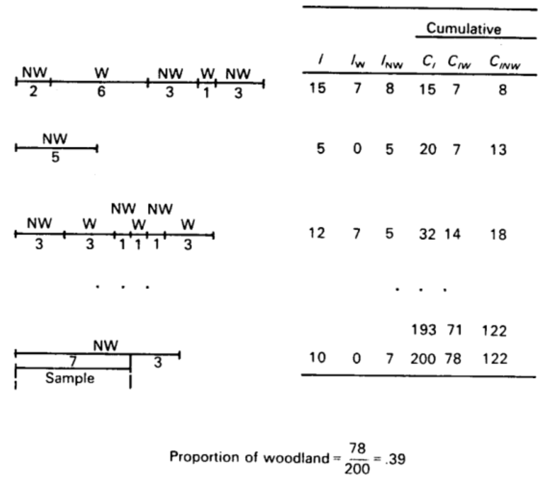


FIGURE 7-16. Estimating the proportion of woodland on a map by using traverses

- Problem with systematic spatial sampling: If the objects under investigation is regularly spaced the systematic sampling may miss it.