LECTURE MATERIALS 1

ENVIRONMENTAL STATISTICS

1. Environmental Statistics: An Introduction

Environmental Statistics

Any models or methods applicable to situations involving uncertainty and variability will be relevant in on guise or another to the study and interpretation of environmental problems and will thus be part of the armoury of *environmental statistics* or *environmetrics***.** Barnett **(2004)**

Measuring the environment is an awesome challange, there are so many things to measure, and at so many times and place. Hunter **(1994)**

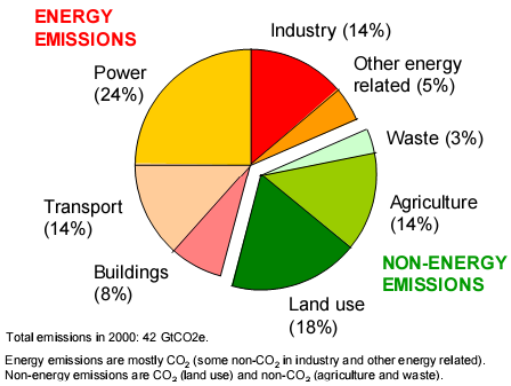
Tomorrow is Too Late

- The average European deposits in a lifetime a monument of waste amounting to about **1000** times body weight; the average North American achieves four times this.
- Sea-floor sediment deposits around the UK average **2000** times of plastic debris per square meter.
- Over their lifetime, each person in the Western world is responsible for carbon dioxide emissions with carbon content on average **3500** times the person's body weight. Harrison **(1992)**

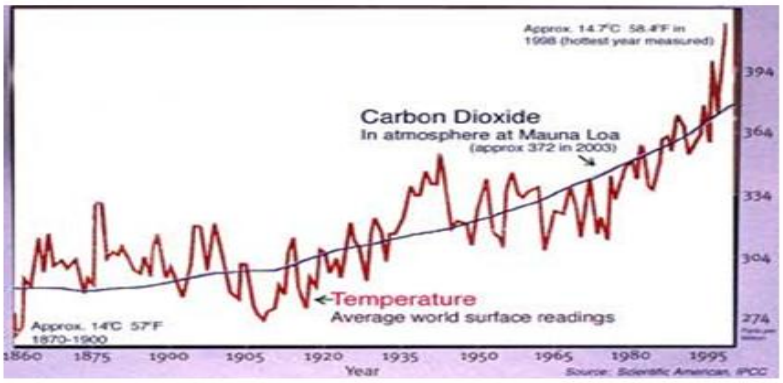
Well-known Environmental Issues

- Acid rain
- Accumulation of greenhouse gases
- Climate change
- Global warming
- Deforestation
- Disposal of nuclear waste products
- Nitrate leaching
- Particulate emissions from disel fuel
- Polluted streams and rivers

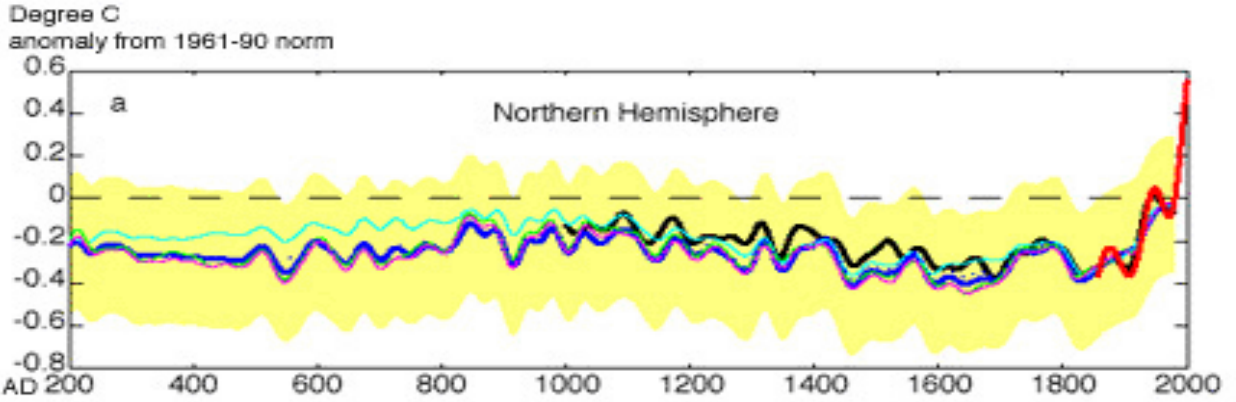
... to name a few...



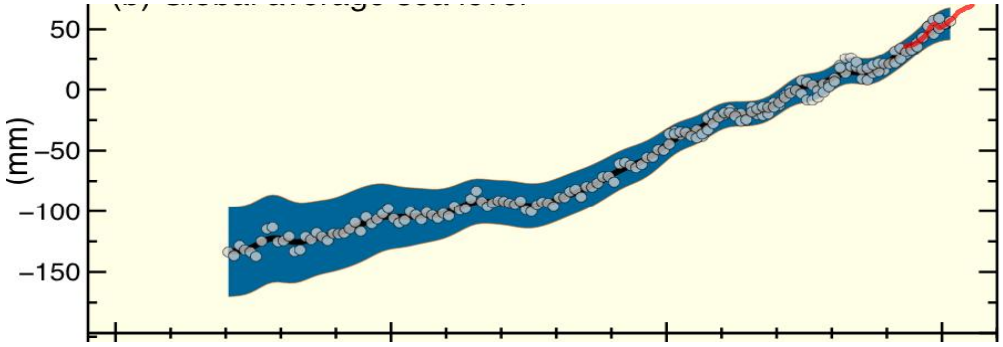
Greenhouse Gasses emitted in 2000, by source (Stern, 2006).



The earth has had highs and lows, droughts and floods, but nothing has been like the past 150 years.



Temperatures AD 200-2000, from proxy temperature indicators and direct measurement (red), showing rise from from long-term cooling trend. Mann & Jones, Geophys. Research Letters, 2003.



2. Extremes, Outliers

Order Statistics and Extremes

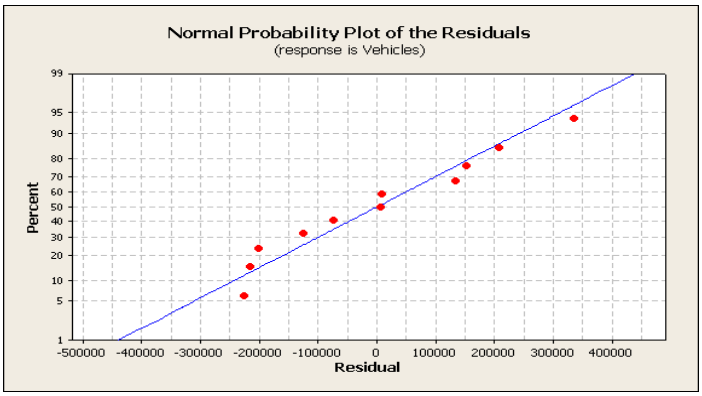
Order statistics are the observations of the random sample, arranged, or ordered, in magnitude from the smallest to the largest.

If $X_1, X_2, ..., X_n$ are observations of a random sample of size *n* from a continuous distribution. Then the random variables $Y_1 < Y_2 < ... < Y_n$ denote the order statistics of the sample.

Order statistics have wide applications in statistics. Most of the measures of location and dispersion such as the five number summary, range, trimmed mean, inter-quartile range are order statistics.

Order Statistics and the Normal Probability Plot

Malaysian road accident data



Distribution of the *r*th order statistics

Let $Y_1 < Y_2 < ... < Y_n$ be the order statistics of n independent observations from a distribution of continuous type with distribution function $F(x)$ and p.d.f. $F'(x) = f(x)$. Then the p.d.f. of the rth order statistics is given by

$$g_r(y) = \frac{n!}{(k-1)!(n-r)!}[F(y)]^{r-1}[1-F(y)]^{n-r} f(y)$$

The largest and the smallest order statistics are known as extremes. It is worth noting that the p.d.f. of the smallest order statistics is

$$g_1(y) = n[1-F(y)]^{n-1} f(y)$$

and the p.d.f. of the largest order statistics is

$$g_n(y) = n[F(y)]^{n-1} f(y)$$

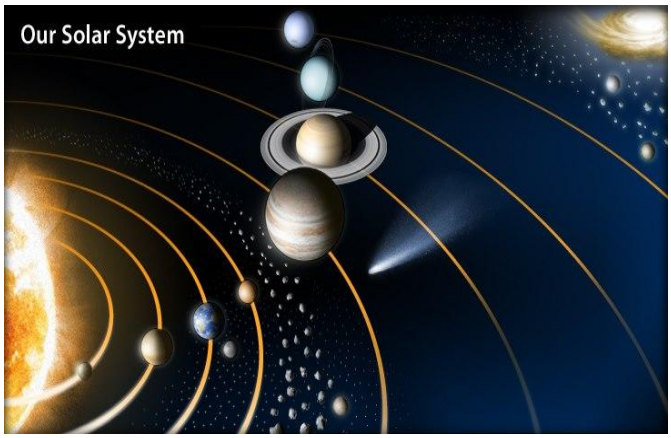
Outliers

'We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.'

– V. Barnett and T. Lewis (1994)

Early History of the Development of Outlier Techniques

The concept of outlier came from Astrophysics even before the formal development of statistics and statistical techniques. The term **'outlier'** was used in astrophysics to distinguish planets which are 'outlying' in our solar system.



Hadlum Versus Hadlum and the Gestation Issue

In 1949, in the case of **Hadlum vs Hadlum**, Major Hadlum appealed against the failure of an earlier petition of divorce. His claim was based on an alleged adultery by Mrs. Hadlum, the evidence for which consisted of the fact that Mrs. Hadlum gave birth to a child which was **349 days** later than when Major Hadlum had left the country to serve the nation during the World War II. The appeal judge **rejected** the appeal.

In other similar cases conflicting views had prevailed. In **Mr. T vs Mrs. T** case also in **1949** the court had ruled that **340 days** was impossible based on the fact that the average gestation period for the human female is **280 days**.

A much earlier case resurfaced at about the same time. In **1921**, **Mr. Gaskil** failed in a petition for divorce on the grounds of adultery based on an absence of **331 days** from home.

Biological Father Versus Statistical Father

Major Hadlum --- **Reject divorce petition: 349 days**

Mr. T --- **Approve divorce petition : 340 days**

Mr. Gaskil --- **Reject divorce petition : 331 days**

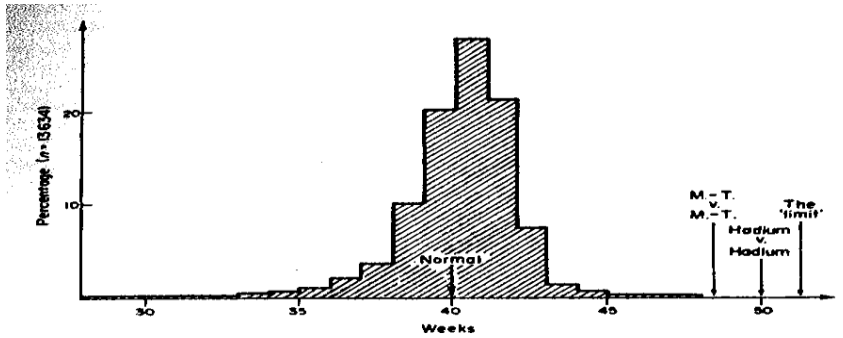


Figure 1.1 Histogram of human gestation periods (reproduced from Barnett, 1978a, by permission of the Royal Statistical Society)

British Court of Justice Decisions

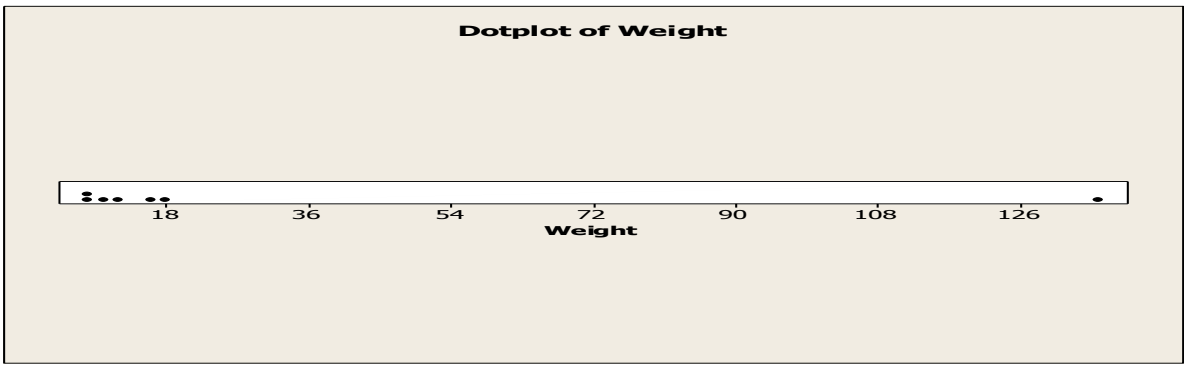
In 1951, the House of Lords had ruled that the limit is 360 days based on a huge survey conducted by the British Medical Association for a sample of 13634 British Births.

Major Hadlum, Mr. T and Mr. Gaskil are all *statistical fathers*.

Outliers Are Empirical Reality

Hampel *et al*. (1986) claim that a routine data set typically contains about 1-10% outliers, and even the highest quality data set can not be guaranteed free of outliers.

Example: Weight of a baby (in lbs): 7.2, 8.6, 10.0, 11.8, **135**, 15.8, 17.6



Consequences of Outliers

One immediate consequence of the presence of outliers is that they may cause apparent non-Normality and the entire classical inferential procedure might breakdown in the presence of outliers.

Summary Statistics of the Baby Weight Data with and without Outliers

Statistic	Without Outlier	With Outlier
Mean	11.83 lbs	29.4 lbs
Standard Deviation	4.11 lbs	46.7 lbs
Range	10.40 lbs	127.8 lbs

Sources of Outliers

Inherent Variability: Natural feature of a population that is uncontrollable.

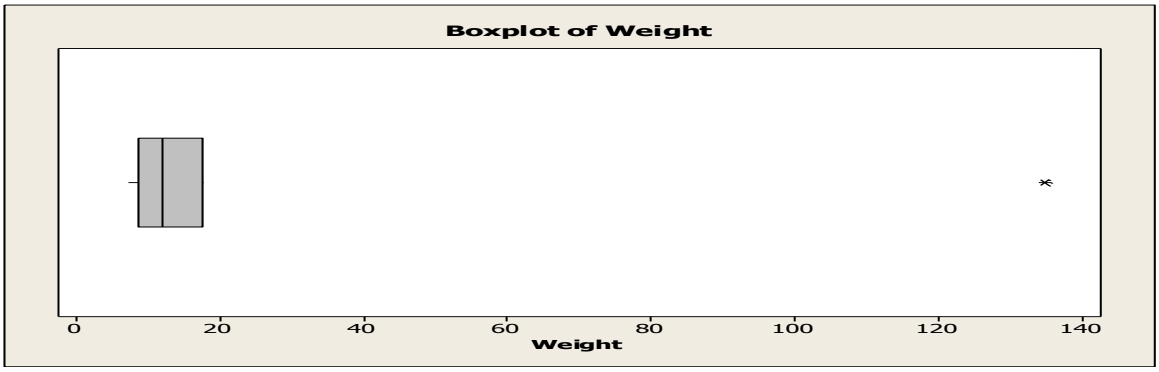
Measurement Error: The rounding of obtained values or mistakes in recording compound measurement error.

Execution Error: Imperfect collection of data. We may inadvertently choose a biased sample or include individuals not truly representative of the population we aimed to sample.

Observations arising from large variation of the inherent type are called outliers, while observations subjected to large measurement error or execution error are termed *spurious observations* (Anscombe, 1960).

Outliers do not inevitably 'perplex' or 'mislead'; they are not necessarily 'bad' or 'erroneous', and the experimenter may be tempted in some situations not to reject an outlier but to welcome it as an indication of some unexpectedly useful industrial treatment or surprisingly successful agricultural variety.

Identification of Outliers



The 'three – sigma' Rule

If we assume a normal distribution, a single value may be considered as an outlier if it falls outside a certain range of the standard deviation.

A traditional measure of the 'outlyingness' of an observation x_i with respect to a sample is the ratio between its distance to the sample mean and the sample SD:

$$t_i = \frac{x_i - \bar{x}}{s}$$

Observations with $| t_i | > 3$ are traditionally deemed as suspicious (the three-sigma rule), based on the fact that they would be very unlikely under normality, since P ($|t| > 3$) = 0.003 for a random variable *t* with a standard normal distribution.

Weight	***t***
7.2	-0.475375
8.6	-0.445396
10	-0.415418
11.8	-0.376874
135	**2.26124**
15.8	-0.291221
17.6	-0.252677

So is there no outlier in this data set???

Masking occurs when we fail to detect the outliers (***false negative***)

Swamping occurs when observations are incorrectly declared as outliers (***false positive***)

Grubbs' Test

Grubbs (1969) proposed a test to detect outliers in a univariate data set. It is based on the assumption of normality. Grubbs' test is also known as the *maximum normed residual test*.

The test statistic is defined as

$$G = \frac{\max | x_i - \bar{x} |}{s}$$

with $\bar{x}$ and s denoting the sample mean and standard deviation respectively.

The Grubbs test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation. For the two-sided test, the hypothesis of no outliers is rejected if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{((\alpha/2n),n-2)}}{n-2+t^2_{((\alpha/2n),n-2)}}}$$

with α denoting the *critical value* of the t distribution with $n - 2$ degrees of freedom and a significance level of α /(2n).

For the weight of baby data we obtain the value of $G = 2.26$. At the 5% level of significance the critical value is 2.02. Thus Grubbs test identifies the case 5 as an outlier.

Grubbs' test detects one outlier at a time. This outlier is expunged from the dataset and the test is iterated until no outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or less since it frequently tags most of the points as outliers. This test may not be effective when swamping occurs in the data.

Dixon's Q-test

The Dixon's Q-test is a very simple test for outliers when we suspect that outliers are extreme observations in the data set.

Q-test is based on the statistical distribution of "subrange ratios" of ordered data samples, drawn from the same normal population. Hence, a normal distribution of data is assumed whenever this test is applied.

The test is very simple and it is applied as follows:

1. The n values comprising the set of observations under examination are arranged in ascending order: $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

2. The statistic Q is a ratio defined as the difference of the suspect value from its nearest one divided by the range of the values. Thus, for testing $x_{(1)}$ or $x_{(n)}$ (as possible outliers) we use the following values:

$$Q = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \quad \text{or} \quad \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$$

3. The obtained Qobs value is compared to a critical Q-value (Qcrit) found in tables.
4. If Qobs > Qcrit, then the suspect value can be characterized as an outlier.

Table for Critical Values of Q

N	Q_{crit} (CL:90%)	Q_{crit} (CL:95%)	Q_{crit} (CL:99%)
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

For the weight of baby data, the value of Q = 0.9186. At the 5% level the critical value is 0.568. Thus this test identifies the case 5 as an outlier.

3. Robust Statistics

Risk in Deleting Outliers

Summary Statistics of the Baby Weight Data with and without Outliers suggests that a simple way to handle outliers is to detect them and remove them from the data set. Deleting an outlier, although better than doing nothing, still poses a number of problems:

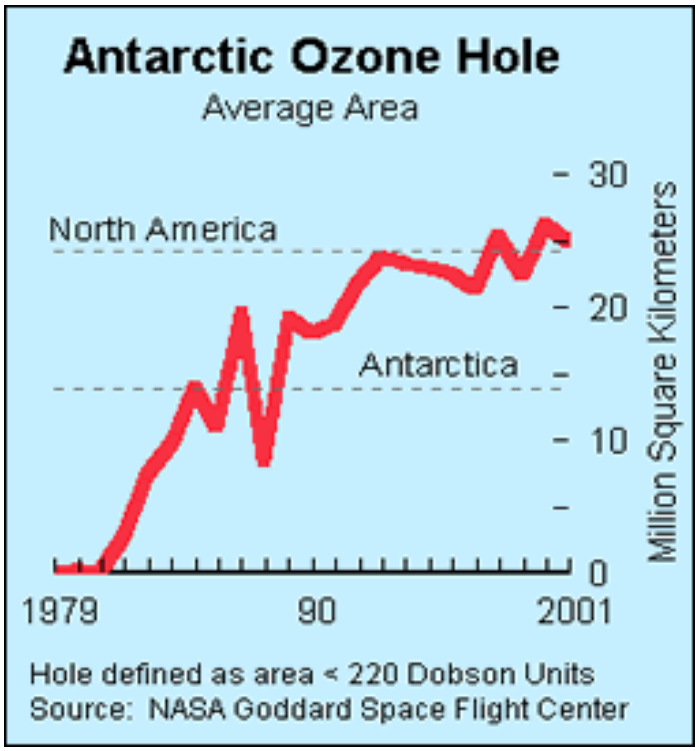
- When is deletion justified? Deletion requires a subjective decision. When is an observation 'outlying enough' to be deleted?
- The user may think that 'an observation is an observation' (i.e., observations should speak of themselves) and hence feel uneasy about deleting them. Sometimes atypical data may be the most informative data and its deletion may outliers.
- Since there is generally some uncertainty as to whether an observation is really atypical, there is a risk of deleting 'good' observations, which results in underestimating data variability.
- Since the results depend on the user's subjective decisions, it is difficult to determine the statistical behaviour of the complete procedure.

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



The word "Robust" literary means something "very strong." So robust statistics are those statistics which do not breakdown easily. The term ***robustness*** signifies ***insensitivity*** to small deviations from the assumption. That means a robust procedure is nearly as efficient as the classical procedure when classical assumptions hold strictly but is considerably more efficient over all when there is a small departure from them. One objective of robust techniques is to ***cope with outliers*** by trying to keep small the effects of their presence. Consequently, we should require ***resistant estimators*** (Tukey, 1977). The analogous term used in the literature: Resistant Statistics

Classical and Robust Approaches to Statistics

Main features:

- Data collected in a broad range of applications frequently contain one or more *atypical observations*, called *outliers*.
- Classical estimates can be very adversely influenced by outliers, even by a single one.
- There exist *robust* parameter estimate that provide satisfactory results when the data contain outliers, as well as when the data are free of them.

Here we introduce several statistics which are robust in the presence of outliers. Median and trimmed mean are robust measures of location. For the measure of dispersion we can use the normalized median absolute deviation (MADN). For a set of data the *Median Absolute Deviation* (MAD) is defined as

$$\text{MAD}(x) = \text{Med}\{|x - \text{Med}(x)|\}$$

To make the MAD comparable to the SD in terms of efficiency, we consider the normalized MAD defined as

$$\text{MADN}\,(x) = \text{MAD}\,(x) / 0.6745$$

Two other well-known dispersion estimates are the *range* defined as

$$R = x_{(n)} - x_{(1)}$$

and the *inter-quartile range* (IQR) defined as

$$\text{IQR}\,(x) = Q_3 - Q_1$$

Both of them are based on order statistics; the former is clearly very sensitive to outliers, while the latter is not.

Summary Statistics of the Baby Weight Data with and without Outliers

Statistic	Without Outlier	With Outlier
Mean	11.83 lbs	29.4 lbs
Median	**10.90 lbs**	**11.80 lbs**
Standard Deviation	4.11 lbs	46.7 lbs
MADN	**4.45 lbs**	**5.93 lbs**
Range	10.40 lbs	127.8 lbs
IQR	**8.0 lbs**	**9.0 lbs**

Robust Outlier Detection Methods

Robust t like Statistic

Let us now use the robust plug-in technique Imon, Midi and Rana (2013) to obtain a robust t-like statistic by replacing mean by median and SD by the normalized median absolute deviation (MADN). Thus the modified statistic becomes

$$t_i' = \frac{x_i - \text{Median}(x)}{\text{MADN}(x)}$$

Observations with $| t_i' | > 3$ are identified as outliers.

Now we compute robust t like statistic as given below.

Weight	**t**	**Robust t**
7.2	-0.475375	-1.77278
8.6	-0.445396	-1.23324
10	-0.415418	-0.693695
11.8	-0.376874	0
135	2.26124	**47.4796**
15.8	-0.291221	1.54154
17.6	-0.252677	2.23524

The above results show that robust t can correctly identify the outlier.

Interquartile Range

The above-mentioned strategies for identifying outliers are probably most appropriate for symmetric unimodal distributions.

If a distribution is skewed, it is recommended to calculate the threshold for outliers from the interquartile distance:

$$Q_1 - 1.5\ \text{IQR} < x_i < Q_3 + 1.5\ \text{IQR}$$

For the weight of the baby data, we obtain

$$Q_1 = 8.6 \qquad Q_3 = 17.6 \qquad \text{IQR} = 9$$

The threshold values for this data set are 0 and 41.1. Hence the case 5 is declared as an outlier.

Hampel's Test

In recent years Hampel (1984)'s test for outliers has become very popular in data mining and knowledge discovery [see Ben-Gal (2005)]

According to this rule an observation x_i is identified as an outlier if

$$x_i - \text{median}(x) > 4.5\ \text{MAD}(x)$$

It is interesting to note that Hampel's test is equivalent to robust t test. Recall that according to the robust t test an observation is identified as an outlier if

$$t_i' = \frac{x_i - \text{Median}(x)}{\text{MADN}(x)} > 3$$

which yields

$$x_i - \text{median(x)} > 3\text{MADN}(x) = 4.4474\ \text{MAD}(x)$$

For the weight of baby data, observation number 5 exceeds this threshold and hence is identified as an outlier.

4. Sampling in the Wild and Population Size Estimation

Conventional Sampling Techniques

- **Simple Random Sampling**
- **Stratified Random Sampling**
- **Systematic Sampling**
- **Cluster Sampling**

- **Non-Probability Sampling: Internet Survey**
- **Hybrid Sampling**

The problem of estimating the finite population size occurs in many branches of statistics. It has an wide application in the analysis of ecological data. In the last fifty years there has been a growing realization of the importance of a sound statistical technique in the analysis of ecological data. Ecologists have also recognized the importance of obtaining data in the field from 'natural' or free-ranging populations rather than 'artificial' or laboratory populations. So often the population changes that occur in the laboratory give little indication as to what happens in the natural state. But the study of natural populations is not easy, since the population size is often not known and we need to estimate the population size before any further analysis.

Well-known problems of this kind are the estimation of the total number of fish in a lake, the estimation of the total number of birds or wild animals in a forest etc. Occasionally it is possible to count all the animals of a particular species in a given area. Seber (1982) presented several examples where animals which congregate in groups could often be photographed and counted later, echo-sounding was used for counting fish, fish which migrate through rivers during part of their life cycles could be counted individually using traps or weirs, radar had been used for estimating bird densities. But in reality it is impossible to count the animals over the whole area because of the disturbances caused or the number of personnel required. In this case a sampling scheme is required.

Several authors had already considered the problem in the past and had suggested different methods of sampling with estimation procedures [see Seber (1982), Boswell *et al*. (1988)]. The basic procedure is to initially draw an object from an urn, color it and put back into the urn and then to draw objects randomly from that urn to recatch the colored object. This approach is usually known as the urn model approach. For the ecological data this approach is equivalent to capture, mark and recapture approach that is popularly known as the C-M-R approach.

Quadrat Sampling

Quadrat samplings is often used in ecological studies. If we wish to count the numbers of one, or of several, species of plant in a meadow to estimate population size or assess biodiversity we might throw a quadrat at random and do our counts within this boundary.

A **quadrat** is usually a square light wood or metal frame of a meter or several meters side. Where it lands defines the search area in which we take appropriate measures of numbers of individual plants, biomass, or extent of ground cover.

Simulation Type Sampling

Recapture Sampling

A wide range of sampling methods are based on the principle of initially '**capturing**' and '**marking**' a sample of the numbers of a finite population and subsequently observing, in a later or separate independent random sample drawn from the same population, how many marked

individuals are obtained. This technique is known as ***recapture sampling*** which is also popularly known as ***capture-recapture sampling*** or ***capture-mark-recapture*** sampling. The sample information is then used to infer characteristics of the overall population, principally its total size.

Points to Remember

- The marking process may be multi-stage, with separate capture and recapture episodes.
- Individuals in the sample can be chosen with or without replacement
- The marking process may sometimes contaminate the population
- Individual may be 'trap shy' (avoid contact) or 'trap happy' (eager for contact, perhaps seeking for food).
- The population may change from capture to recapture due to births, deaths or inward or outward transfer.

Marking Processes

- Rings on the legs birds
- Small radio transmitters inserted under the skin of animals
- Radioisotope marking
- Nicks or cuts on the fins of fish
- Color marking of skin or fur
- Cutting patterns in the bark of trees
- Tying a colored plastic or material marker to an individual

Estimation of Population Size: The Peterson and Chapman Estimators

Let us suppose that an initial random sample of size n is chosen from a population of size N. Each of the sample is marked and the sample is then returned to the population. A second sample of size m is then taken and turns out to contain r of the originally marked individuals. Then the Peterson estimator of the population size is given by

$$\hat{N} = \frac{nm}{r}$$

An approximate unbiased estimator of its variance is given by

$$V\left(\hat{N}\right) = \frac{nm(n-r)(m-r)}{r^3}$$

In a similar situation the Chapman estimator of the population size is given by

$$\hat{N} = \frac{(n+1)(m+1)}{(r+1)} - 1$$

with an approximate unbiased estimator of variance

$$V(\hat{N}) = \frac{(n+1)(m+1)(n-r)(m-r)}{(r+1)^2(r+2)}$$

Urn Model Approach

The simplest of all of the methods of population size estimation. Initially one population unit is drawn at random, colored and is sent back into the target population. In the next step one unit is drawn at random. If the drawn unit is the colored one then the sampling is stopped, otherwise it is colored and is sent back into the population. This procedure is repeated until a colored unit is drawn for the first time.

Let S denote the effective number of trials required to get a colored unit for the first time. Then the estimate of the population size is obtained by

$$\hat{N}_{UE} = \binom{s+1}{2}.$$

Alam, Imon and Sinha (2006) suggest a very simple approximation of the maximum likelihood estimator of finite population size given by

$$\hat{N}_{MLE} \approx \frac{s^2}{2} + \frac{5s}{32} \quad \text{for all } s$$

The following table offers a comparison between the unbiased and the ML estimate of population size. (Source: Alam, M. M., Imon, A. H. M. R. and Sinha, B. K (2006). Maximum Likelihood Estimation of a Finite Population Size, *Journal of Statistical Theory and Applications*, Vol. 5, No. 3, pp. 306 – 311.)

Table. Unbiased and ML estimate of finite population size for different no. of trials

$S = s$	$\hat{N}_{UE}$	$\hat{N}_{ML}$	$S = s$	$\hat{N}_{UE}$	$\hat{N}_{ML}$
1	1	1	26	351	342
2	3	2	27	378	369
3	6	5	28	406	396
4	10	9	29	435	425
5	15	13	30	465	455
6	21	19	31	496	485
7	28	26	32	528	517
8	36	33	33	561	550
9	45	42	34	595	583
10	55	52	35	630	618
11	66	62	36	666	654
12	78	74	37	703	690

13	91	87	38	741	728
14	105	100	39	780	767
15	120	115	40	820	806
16	136	130	41	861	847
17	153	147	42	903	889
18	171	165	43	946	931
19	190	183	44	990	975
20	210	203	45	1035	1020
21	231	224	46	1081	1065
22	253	245	47	1128	1112
23	276	268	48	1176	1160
24	300	292	49	1225	1208
25	325	316	50	1275	1258

Adaptive Sampling

In adaptive sampling we often collect more samples from a location where there is higher concentration of objects instead of a simple random sampling.