LECTURE MATERIALS 2

ENVIRONMENTAL STATISTICS

5. Inaccessible and Sensitive Data

1.GPA
2.Weight
3. Money besides salary
4. Income
5.Caloria
6. sexual habit
7.Alchohol for minor
8. women--abortion or not
9. political
10. drunk driving
11.drug habbit
12. stole before or not
13. married with kid.
14. Fin Assistance

Application of Conventional Sampling Techniques for Sensitive Data

Estimation of Population Total

Let us suppose that the observation from the first respondent is x_1 but for some reason he/she is not willing to disclose that information. But he/she will not mind to pass the information $a + x_1$ where $a \geq 0$ is a secret number which nobody except the first respondent knows. This secret number is also known as hidden, base or seed number. The second respondent will have absolutely no idea about x_1 since he/she does not know the value a. He/she then add his/her observation x_2 with $a + x_1$ and pass it to the third respondent. The third respondent will only see the value $a + x_1 + x_2$ but will have no idea about the individual x_1 or x_2. This process is continued till the last respondent includes his/her information and the quantity

$$a + x_1 + x_2 + \ldots + x_n = a + \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} u_i \text{ (say)}$$

is obtained. The information $\sum_{i=1}^{n} u_i$ is then sent to the first respondent for the exclusion of the secret value a. The first respondent then subtract number a from $\sum_{i=1}^{n} u_i$ and pass the final result to the experimenter. Thus the final result

$$\hat{X} = \sum_{i=1}^{n} u_i - a = \sum_{i=1}^{n} x_i$$

which is the estimate of the population total obtained by the conventional simple random sampling

Estimation of Population Mean

For the estimation of population mean for sensitive data we follow exactly the same procedure described above. After obtaining the value $\sum_{i=1}^{n} u_i - a$ from the first respondent we estimate the mean by

$$\hat{\mu} = \frac{\sum_{i=1}^{n} u_i - a}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

which is again the estimate of the population mean obtained by the simple random sampling.

Estimation of Population Proportion

For the estimation of population proportion for sensitive data we follow similar approach described above. Since we are estimating proportion, the values of x_1 can take only two numbers; 0 for one group and 1 for the other group. It is worth mentioning that the secret number *a* should not take value 0 because if the first respondent does not have the characteristic under our study, the quantity $a + x_1$ will be zero and then the second respondent would definitely realize that the first respondent did not have that characteristic. If we denote the total count by *Y*, the *n*th respondent will pass the value $W = Y + a$ to the first respondent. The experimenter will get back the number $W - a$ from the first correspondent and the estimate of the population proportion is then estimated by

$$\hat{p} = \frac{W - a}{n} = \frac{Y}{n} = P$$

which is the estimate of the population proportion obtained by the simple random sampling.

Nonconventional Sampling Techniques

Network Sampling: Network sampling utilizes a "word of mouth" approach of acquiring participants. Those who are originally recruited suggest further participants. This method allows researchers to access populations that are not easily identifiable, are small in number, private, poorly organized or socially marginalized. Examples of such populations would be sexual minorities, drug users, etc.

The advantage of network sampling is that these hard-to-reach populations are penetrated and recruitment is fairly convenient and inexpensive for the researcher. Most research methods experts find that network sampling is just as effective as other, more random methods and rarely leads to validity or reliability errors.

There is an essential need to constantly monitor the environment for changes in level of pollutants, industrial by-products, etc. This process can take the form of regular sampling of a fixed set of sites, often arranged roughly on a grid or network. We employ network sampling in this regard.

Encounter Sampling: This is a data collection procedure in which population units are included in the sample as they are detected or encountered. We have to consider encounter sampling when we have to take what is to hand or we may have to ensure optimum use of scarce resources either by economizing in our number of observations or by exploiting any form of circumstantial information that is available.

Encountered Data

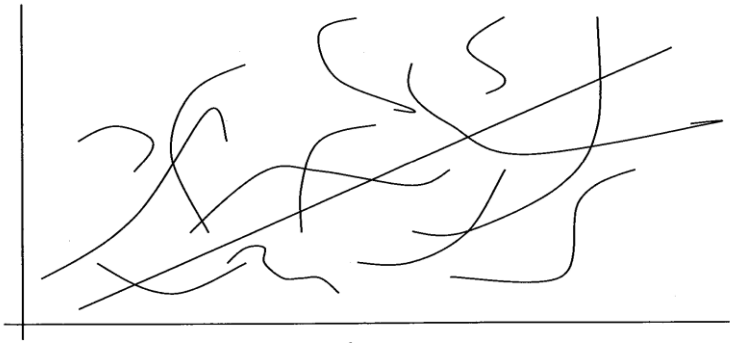
A data set is known as encountered data when the investigator goes into the field, observes and records what he observes… what he/she encounters. The long-established data collection techniques need observations to be taken at random and under prescribed circumstances. This is often not possible with environmental problems – we have instead to make do with what forms or and limited numbers of observations can be obtained, and on the occasions and at the places they happen to arise.

For example, climatological variables observed over time, and especially in the past, have to be limited to what was collected by the meteorological station; inundations and tornados occur when they occur! Measured pollution levels tend to be taken and published selectively, for example when site visits are made, perhaps because of the suspicion that levels have become rather high. Accessibility is an important factor here; measurements can only be taken when they are allowed to be taken, to the extent to which they can be afforded, when equipment is available, when they happen to have been taken, and so on.

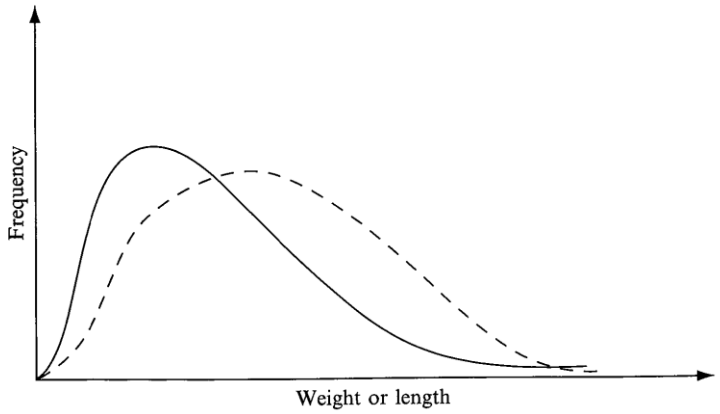
Length-Based or Size-Based Sampling and Weighted Distributions

If we sample fish in a pond by catching them in a net, there will be encounter bias (more usually called size bias). This is because the mesh size will have the effect of lowering the incidence of the smaller fish in the catch- some will slip through the net.

If we were to sample harmful industrial fibers (in monitoring adverse health effects) by examining fibers on a plane sticky surface by line-intercept methods, the similar problem may arise. In this case our data would consist of the lengths of fibers crossed by the intercept line as shown below.



Our interest will be in the distribution of sizes, but the sampling methods just described are clearly likely to produce seriously biased results. Here we are bound to obtain what are known as length-biased or size-biased samples, and statistical inference drawn from such samples will be seriously flawed because they relate to distribution of measured sizes, not to the population at large (as shown in the following figure), which will our real interest. Thus we will typically overestimate the mean both in the fish and in the fiber examples, possibly to a serious extent.



Weighted Distribution Methods

Suppose X is nonnegative random variable with mean μ and variance σ^2, but what we actually sample is a random variable X^*. A special but popular case of the size-biased distribution has the p.d.f.

$$f^*(x) = xf(x)/\mu$$

The variable actually sampled has expected value

$$E^*(X^*) = \int [x^2 f(x)/\mu] dx = \mu\left(1+\frac{\sigma^2}{\mu^2}\right)$$

So if we take a random sample of size n, then the sample mean of the observed data $\bar{x}^*$ is biased upward by a factor $\left(1+\frac{\sigma^2}{\mu^2}\right)$.

Here the problem is that we do not know the true values of μ and σ^2. However, the statistic

$\frac{\bar{x}^* \sum_{i=1}^{n} 1/x_i^*}{n}$ provides an intuitively appealing estimate of the bias factor $\left(1+\frac{\sigma^2}{\mu^2}\right)$.

Example: Consider the following 24 determinations of the copper content in wholemeal flower (in parts per million)

.

2.2	2.2	2.4	2.5	2.7	2.8	2.4	2.9
3.03	3.03	3.1	3.37	3.4	3.4	3.4	3.5
3.6	3.7	3.7	3.7	3.7	3.77	5.28	28.95

With Outlier		**Without Outlier**	
X^*	$1/X^*$	X^*	$1/X^*$
2.20	0.454545	2.20	0.454545
3.03	0.330033	3.03	0.330033
3.60	0.277778	3.60	0.277778
2.20	0.454545	2.20	0.454545
3.03	0.330033	3.03	0.330033
3.70	0.270270	3.70	0.270270
2.40	0.416667	2.40	0.416667
3.10	0.322581	3.10	0.322581
3.70	0.270270	3.70	0.270270
2.50	0.400000	2.50	0.400000
3.37	0.296736	3.37	0.296736
3.70	0.270270	3.70	0.270270
2.70	0.370370	2.70	0.370370
3.40	0.294118	3.40	0.294118
3.70	0.270270	3.70	0.270270
2.80	0.357143	2.80	0.357143
3.40	0.294118	3.40	0.294118
3.77	0.265252	3.77	0.265252
2.40	0.416667	2.40	0.416667
3.40	0.294118	3.40	0.294118
5.28	0.189394	5.28	0.189394
2.90	0.344828	2.90	0.344828
3.50	0.285714	3.50	0.285714
28.95	0.034542		
$\bar{x}^* = 4.28$		$\bar{x}^* = 3.208$	

	Bias Factor		**Corrected Mean**	
	Mean Based	**Median Based**	**Mean Based**	**Median Based**
With Outlier	1.33295	0.998452	3.21092	3.38524
Without Outlier	1.04260	0.999879	3.07692	3.37041

Example: If X has a Poisson distribution, then

$$f^*(x) \propto \frac{e^{-\mu}\mu^{X^8}}{[\mu(x^*-1)!]} = \frac{e^{-\mu}\mu^{X^*-1}}{(x^*-1)!}$$

so that $X^* - 1$ has a Poisson distribution with mean μ. Since $\mu = \sigma^2$, so the bias factor becomes $\left(1+\frac{1}{\mu}\right)$. Hence $\bar{x}^*$ will be the unbiased estimator of $\mu + 1$ and thus $\bar{x}^* - 1$ will be the unbiased estimator of μ.

Random			**Encounter**		
#Defective Teeth	# of children	Total	#Defective Teeth	# of children	Total
0	872	0	0	151	0
1	82	82	1	178	178
2	33	66	2	127	254
3	7	21	3	25	75
4	4	16	4	11	44
5	1	5	5	5	25
6	1	6	6	3	18
	1000	196		500	594

For the random sample the mean of the number of defective teeth of children is 0.196. For the encounter sample the mean is 1.188. The standard literature tells us that the number of defective teeth of children follows a Poisson distribution. Hence after the bias correction the mean of the number of defective teeth of children is 0.188.

Many other weighted distribution methods have been studied and used. For instance, in the fish example with a square mesh of size x_0, the weight function is more likely a truncation and we would have

$$g(x) = \begin{cases} 0 & x \le x_0 \\ 1 & x > x_0 \end{cases}$$

So now

$$f^*(x) = \frac{f(x)}{\int_{x_0}^{\infty} f(x)dx}$$

However, since the parameter is usually unknown this will often not be easy to handle. In one case it is straight forward.

Example: Consider sampling from an exponential distribution with $f(x) = \frac{1}{\theta}e^{-x/\theta}$, $x \ge 0$

Then

$$f^*(x) = \frac{\frac{1}{\theta}e^{-x/\theta}}{e^{-x_0/\theta}} = \frac{1}{\theta}e^{-(x-x_0)/\theta},\ x > x_0$$

So $X^* - x_0$ is exponential with parameter θ, i.e., $E(X^* - x_0) = \theta$. The bias is just (the mesh size) so we use $\bar{x}^* - x_0$ for estimating θ.

Example: The following table gives square mesh of 15 fishes in a pond (in inch)

387	275	228	479	381
301	149	362	366	459
221	73	354	88	478

From the above table we obtain the average mesh of fish as 306.7 sq inches. If the truncation occurs below 36 sq inches, then the bias corrected average mesh of the fishes is 270.7 sq inches.

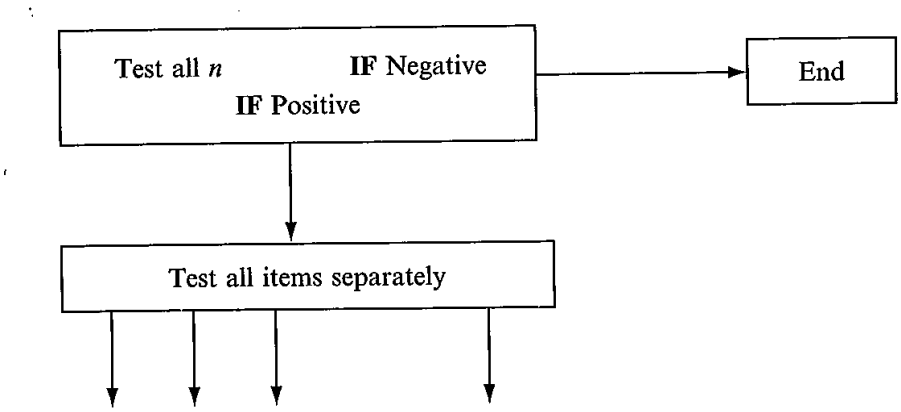
Composite Sampling

Often we need to identify those members of as population who possess some rare characteristic or condition Sometimes the condition is of a 'sensitive' form, and individuals may loath to reveal it. Alternatively it may be costly or difficult to assess each member separately.

One possibility might be to obtain material or information from a large group of individuals, to mix it all together and to make a single assessment for the group as a whole. This assessment will reveal the condition if any one of the group has the condition. If it does not show up in our single test we know that all our members are free of that condition. A single test may clear 1000 individuals!

This is the principle behind what is known as composite sampling. It is also known as aggregate sampling or grab sampling. Of course, our composite sample might show the condition to be present. Then we know nothing about which, or how many, individuals are affected. But that is another matter that we will discuss later.

Early examples of group testing were concerned with the prevalence of insects carrying a plant virus and of testing US servicemen for syphilis in the Second World War. The material collected from each member of a sample is pooled, and a single test is carried out to see if the condition is present or absent; for example, blood samples of patients might be mixed together and tested for the presence of the HIV virus.



Full Retesting

Applications of composite sampling cover a broad range, from testing for presence of disease to examining if materials fail to reach safety limits. Specific examples include: remedial clean-up of contaminated soil, geostatistical sampling, examining foliage and other biological materials, screening of dangerous chemicals, groundwater monitoring, and air quality.

Attribute Sampling

Suppose a population has a proportion p with characteristic A and we take a random sample of size n, but, instead of observing the individual values separately, we test the overall sample (in composite form) once only for the presence of the characteristic A in at least one of the sample members. Then

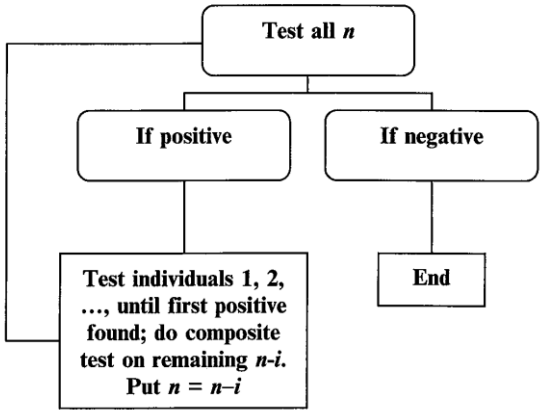
$$P(A \text{ encountered}) = 1-(1-p)^n .$$

If we do not find A, we conclude that no members of the sample of size n have the characteristic.

If we find A, and we need to identify precisely which sample members have the characteristic, we must examine the sample in more detail. The most obvious approach is to retest each sample separately.

Note that it requires some care. If we have used all the sample material for the composite test, we would not subsequently examine each individuals separately without resampling. It would be more prudent, and this is common practice, to use only some of the material in the composite test and to retain some from each individual (so-called ***audit sample***s) for later use if necessary.

In the full retesting approach we will need either one test (if negative) or $n + 1$ tests (if positive) to identify precisely which sample members are affected. We observe that in general we need on average $(n+1)-n(1-p)^n$ tests. So if $p = 0.0005$ and $n = 20$, just 1.2 tests are required on average.

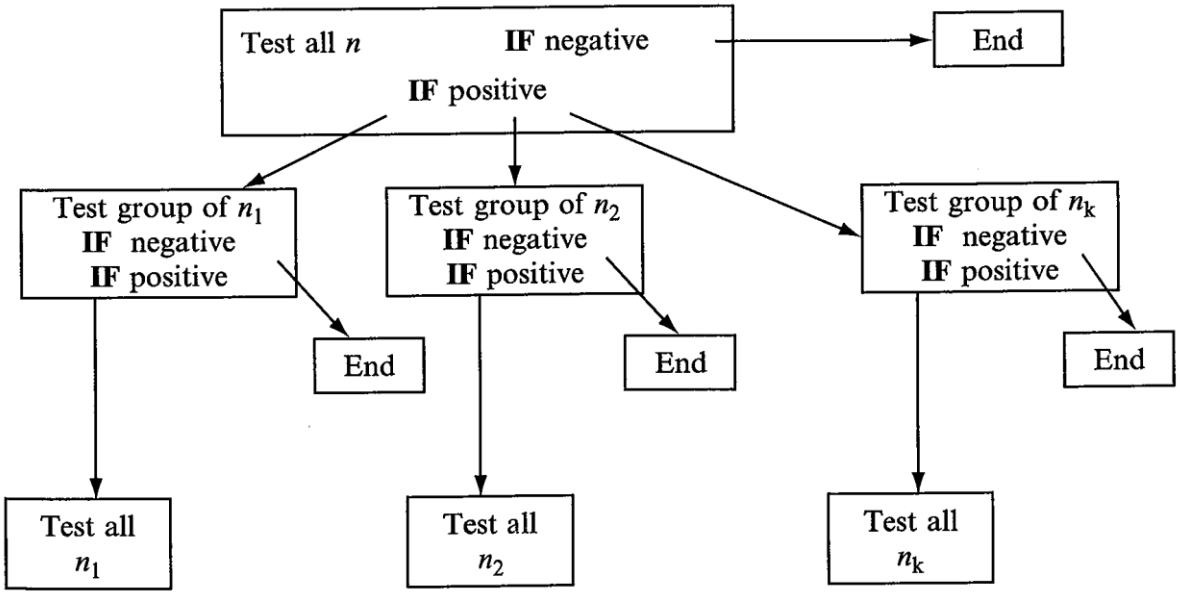


Sudden Death Retesting

If the initial test of the composite sample shows that A is present then various other strategies can be taken to identify the affected samples. We can employ ***group retesting*** or ***cascading*** approach in this regard. Instead of testing each individual we might retest in composite subsamples. A version of this is known as ***sudden death retesting***. Here following a positive first test, we would test the individuals one at a time until we find the first affected individual and then conduct a composite test

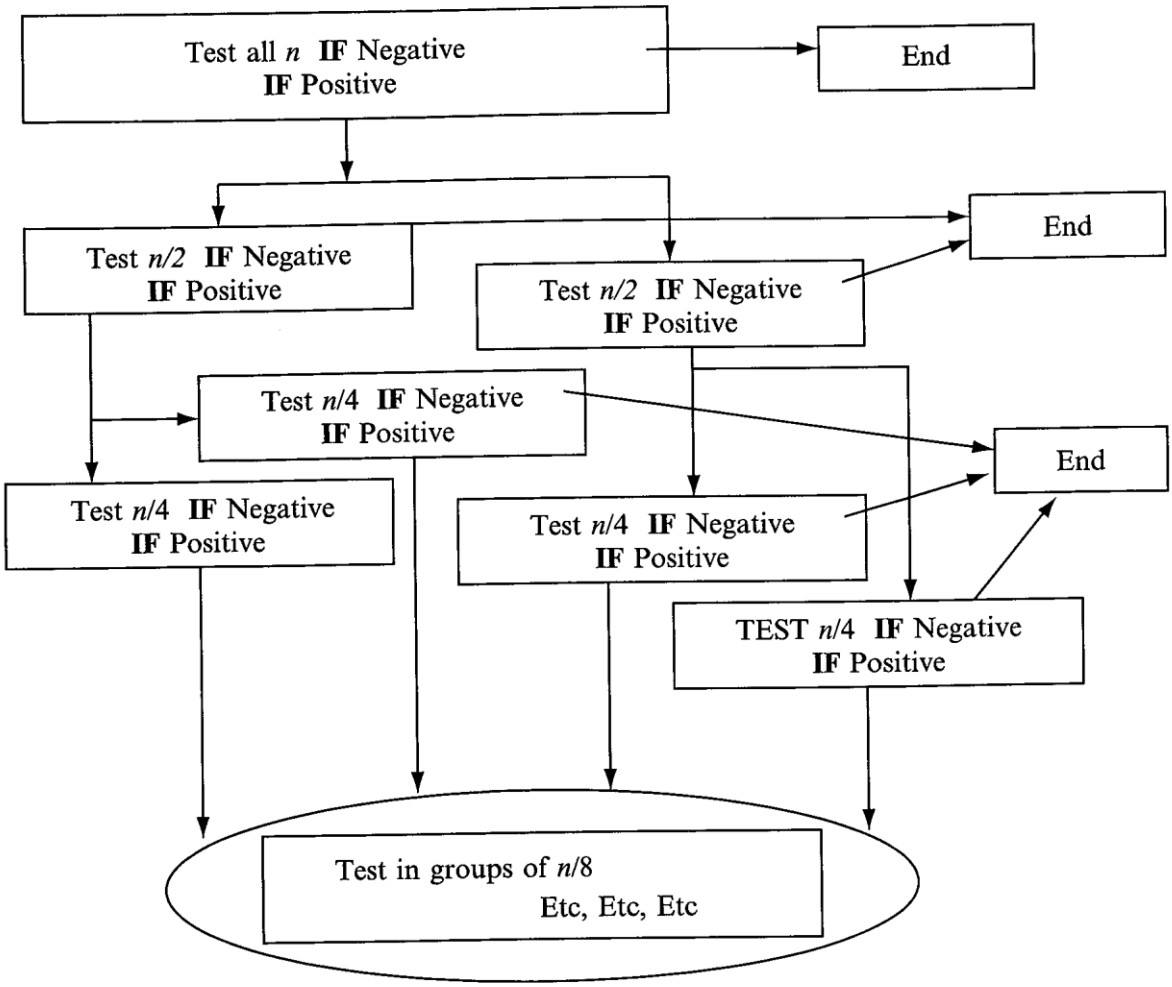
on the remainder. If the composite test is negative we stop the whole process and say we have only one affected sample. If the composite test is positive we repeat the process and so on.

In group retesting we divide the sample group of size n into k subgroups, $n_1, n_2, ..., n_k$ if the first overall composite test is positive. Each of the subgroups is treated as a second-stage composite sample. Each subgroup is then tested as for full retesting and the process terminates.



Group Retesting

A special modification of group retesting is known as cascading where we adopt a hierarchical approach, dividing each positive group or subgroup into two parts and continue testing until all positive samples have been identified.



Cascading

Example: On an urban site previously used for chemical processing, 128 soil samples are chosen from different locations to test for the presence of a particularly noxious substance.

If we opt for full retesting we know that on average we must carry out 129 – 128 $((1-p)^{128}$ tests.

Under cascading it is more difficult to calculate the average number. If only one sample is contaminated we need exactly 15 tests. If all samples are contaminated we need $2^8 - 1 = 255$ tests.

Which alternative is more economical depend on the value of *p*. If *p* is small sudden death is perhaps the best choice. Otherwise group retesting or cascading should be used.

For estimating the value of *p*, we could test *m* composite samples. Suppose *r* of them exhibit characteristic *A*, then *r* is the binomial B[*m*, $1-(1-p)^n$] so that since *r/m* is the MLE of $1-(1-p)^n$, an estimator of *p* is provided by

$$p^* = 1-(1-r/m)^{1/n}$$

Continuous Variables

A modified composite sampling scheme can be considered if X is a continuous random variable. For example, X can measure the pollution levels in a river, and we want to know if any observed x_i in a sample of size n are illegally high values above some control value or standard, x_H. For a composite sample of size n, we compute $\bar{x}$.

If $\bar{x} < x_H/n$, we declare all observations to be satisfactory.

If $\bar{x} \geq x_H/n$, we would need to retest all observations (or smaller composite subsamples).

This procedure is known as '**rule of *n***' composite sampling procedure.

Example: Consider the following 24 determinations of the copper content in wholemeal flower (in parts per million).

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
2.2	2.2	2.4	2.5	2.7	2.8	2.4	2.9
3.03	3.03	3.1	3.37	3.4	3.4	3.4	3.5
3.6	3.7	3.7	3.7	3.7	3.77	5.28	28.95

If the standard copper content level is 5.00 (in parts per million) we can use all previously discussed methods to find the contaminated sample.

Full Retesting	
Observation	Copper Content
1	28.95
2	13.03
3	3.6
4	3.28
5	2.2
6	3.03
7	3.7
8	3.77
9	2.4
10	3.1
11	3.7
12	3.37
13	2.5
14	2.4
15	3.7
16	2.7
	Mean =5.339375

Sudden Death		
Observation	Copper Content	
1	**28.95**	
2	13.03	13.03
3	3.6	3.6
4	3.28	3.28
5	2.2	2.2
6	3.03	3.03
7	3.7	3.7
8	3.77	3.77
9	2.4	2.4
10	3.1	3.1
11	3.7	3.7
12	3.37	3.37
13	2.5	2.5
14	2.4	2.4
15	3.7	3.7
16	2.7	2.7
Mean =	**5.339375**	**3.765333**

Cascading											
Copper Content	Observation	Copper Content	Observation	Copper Content		Observation	Copper Content		Observation	Copper Content
28.95	1	28.95	1	28.95	Mean =	1	28.95	Mean =	**1**	**28.95**
13.03	2	13.03	2	13.03	**12.215**	2	13.03	**20.99**		
3.6	3	3.6	3	3.6					Observation	Copper Content
3.28	4	3.28	4	3.28		Observation	Copper Content		**2**	**13.03**
2.2	5	2.2				3	3.6	Mean =		
3.03	6	3.03	Observation	Copper Content		4	3.28	**3.44**		
3.7	7	3.7	5	2.2	Mean =					
3.77	8	3.77	6	3.03	**3.175**					
2.4	Mean =	**7.695**	7	3.7						
3.1			8	3.77						
3.7	Observation	Copper Content								
3.37	9	2.4								
2.5	10	3.1								
2.4	11	3.7								
3.7	12	3.37								
2.7	13	2.5								
5.339375	14	2.4								
	15	3.7								
	16	2.7								
	Mean =	**2.98375**								

Ranked-Set Sampling

In many areas of environmental risk such as radiation (soil contamination, disease clusters, air-borne hazard) or pollution (water contamination, nitrate leaching, root disease of crops) we commonly find that the taking of measurement can involve substantial scientific processing of materials and correspondingly high attendant cost. ***Ranked-set sampling*** is often used to draw statistical inference as expeditiously as possible with regard to containing the sample costs.

A simple example of the problem arises even when we wish to estimate as basic a quantity as a population mean. It could operate in this way. If we want a sample of size 5 we would chose 5 sites at random, but rather than measuring pollution at each of them we would take the largest pollution

level. We then repeat the process by selecting a second random set of five sites and measure the second largest pollution level amongst these, and so on, until we get the lowest pollution level in the final random set of five sets. The resulting ranked-set sample of size 5 is then used for the estimation of the mean. Such an approach can be used to estimate a measure of dispersion, a quantile or even to carry out a test of significance, or to fit a regression model.

The ranked set sampling approach can be described in the following way. We consider a set of n observations of a random variable X. These would yield observations in the form

x_{11}	x_{21}	x_{31}		x_{n-11}	x_{n1}
x_{12}	x_{22}	x_{32}		x_{n-12}	x_{n2}
x_{13}	x_{23}	x_{33}		x_{n-13}	x_{n3}
x_{1n-1}	x_{2n-1}	x_{3n-1}		x_{n-1n-1}	x_{nn-1}
x_{1n}	x_{2n}	x_{3n}		x_{n-1n}	x_{nn}

Instead of considering all observations we would consider only one measured observation in each subsample, the ith ordered value in the ith sample. The ranked-set sample is then obtained as the diagonal elements of the following table, i.e. $\boldsymbol{x_{1(1)}}, \boldsymbol{x_{2(2)}}, \ldots, \boldsymbol{x_{n(n)}}$.

$\boldsymbol{x_{1(1)}}$	$x_{2(1)}$	$x_{3(1)}$		$x_{n-1(1)}$	$x_{n(1)}$
$x_{1(2)}$	$\boldsymbol{x_{2(2)}}$	$x_{3(2)}$		$x_{n-1(2)}$	$x_{n(2)}$
$x_{1(3)}$	$x_{2(3)}$	$\boldsymbol{x_{3(3)}}$		$x_{n-1(3)}$	$x_{n(3)}$
$x_{1(n-1)}$	$x_{2(n-1)}$	$x_{3(n-1)}$		$\boldsymbol{x_{n-1(n-1)}}$	$x_{n(n-1)}$
$x_{1(n)}$	$x_{2(n)}$	$x_{3(n)}$		$x_{n-1(n)}$	$\boldsymbol{x_{n(n)}}$

Then the ranked-set sample mean is defined as

$$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_{i(i)}$$

It is easy to show that $\bar{\bar{x}}$ is an unbiased estimator of the population mean μ and

$$\mathrm{Var}(\bar{\bar{x}}) \leq \mathrm{Var}(\bar{x})$$

where $\bar{x}$ is the traditional sample mean of all n^2 observations.

Example: The following table gives square mesh of 16 fishes in a pond (in inch)

587	149	479	381
301	73	366	459
221	228	254	478
275	462	88	65

For this data the ranked set sample is

221	73	88	65
275	**149**	254	381
301	228	**366**	459
587	462	479	**478**

Descriptive Statistics: Ranked Set

```
Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
Ranked Set  4   0  303.5     73.6  147.2    149.0  167.0   293.5  450.0

Variable    Maximum
Ranked Set    478.0
```

Descriptive Statistics: Original Set

```
Variable       N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
Original Set  16   0  304.1     39.8  159.2     65.0  167.0   330.0  441.8

Variable      Maximum
Original Set    587.0
```

6. Generalized Linear Models

A wide range of non-linear models can be accommodated under the label ***generalized linear mode***l. One popular model of this kind often used in environmental studies is the ***dose-response model***. In this model a particular level of exposure (dose) of a stimulus may cause an effect in the response of an affected subject. The stimulus may be environmentally encountered (as in the level of sulphur dioxide in the air) or environmentally administered (as in the dose given to plants to kill of infestation or to a patient to ease a condition).

The response may be quantitative (as in the effect on a biomedical blood measure of a patient), or qualitative (in healing the condition or killing the insects). If it is qualitative, we may have what is known as a ***quantal response model***. Such models are widely employed in environmental, epidemiological and toxicological studies.

Example: Suppose we consider the effects of applying different dose levels (in appropriate units) of a trial insecticide for possible control of an environmentally undesirable form of infestation and obtain the following data on number of insects treated and killed at different application levels of the insecticide.

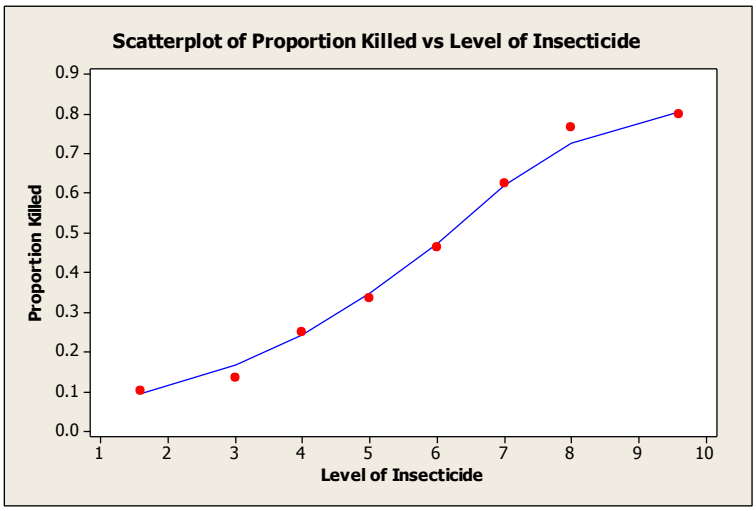
Level of Insecticide	Insects Treated	Insects Killed	Proportion Killed
1.6	10	1	0.100000
3.0	15	2	0.133333
4.0	12	3	0.250000
5.0	15	5	0.333333
6.0	13	6	0.461538
7.0	8	5	0.625000
8.0	17	13	0.764706

9.6	10	8	0.800000

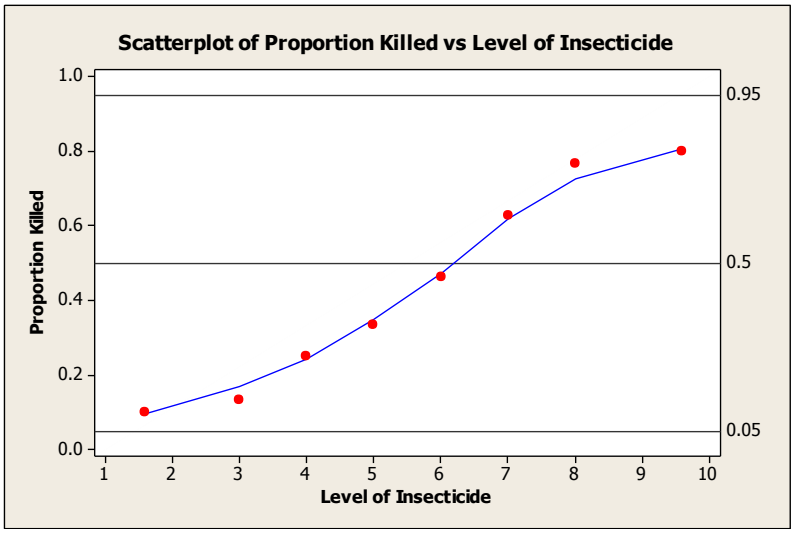
Clearly the proportion killed tends to increase with the level (dose) of the applied insecticide, but it seems to do so in a non-linear way.

Toxicology Concerns

Dose-response relationships are of particular interest in toxicology, where we wish to examine how effective or toxic is the influence of a certain stimulus on the behavior of a response variable. The subject may be human, animal or plant, a community of such beings or even the entire ecosystem.



Dose-response effects are often summarized by certain features of the dose-response relationship. Commonly employed measures include those known as the median lethal dose (LD) LD_{50} level, the LD_5 level and the LD_{95} level: the doses that kill 50%, 5% and 95% of the population, respectively.

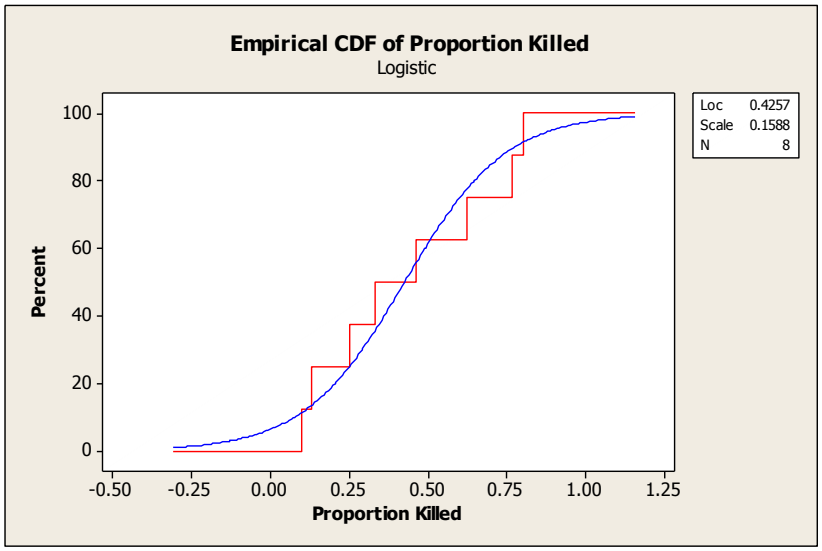


Depending on application the various measures LD_{50}, LD_5, LD_{95} may be of more or less interest. For example, if we are testing a new insecticide we will want to be effective in controlling the offending agent and we will probably concentrate on the LD_{50} to reflect that the agent is under control. If we are concerned about the effect of an environmental pollutant on human health, we will want this effect to be low and will focus on the LD_5 or even lower effect measures such as the NOEL (***no obvious effect level***) or the NOAEL (***no obvious adverse effect level***).

An alternate and related approach applied to the dose-response data across environmental, epidemiological and general medical interests is that of ***benchmark analysis***. The ***benchmark dose*** (BMD) or benchmark level is that dose or exposure level which yields a specific increase in risk compared with incidence in an unexposed population. Such increase is called ***benchmark risk*** (BMR). Benchmark analysis proceeds by specifying the BMR and applying dose-response analysis to estimate the BMD in the form of statistically inferred lower bound. The assumed dose-response model often takes the form of a ***binary quantal response model***- where one of two specific qualitative outcomes must arise at any dose for any individual.

Quantal Response

Let us define a model for the quantal response data. Morgan (1992) proposed that the response variable Y should be a binomial, $Y \sim B[n, p(x)]$, with mean $np(x)$ depending on the level or dose x. We will need to assume that $p = F(x)$, where $F(x)$ is a distribution function. This is often referred to as the distribution function of the tolerance distribution. Thus $F(x)$ takes a general shape as shown below. It is monotonic non-decreasing in x and ranges from 0 to 1.



One form commonly used is the ***normal model*** or ***probit model*** defined as

$$F_1(x) = \Phi(\alpha + \beta x)$$

where $\Phi(z)$ is the distribution function of the standard normal distribution. Another is the logistic model as defined in Chapter 7. In the dose-response study we often consider the ln (natural logarithm) dose and the model becomes

$$F_2(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

For the logit case, it is common to use the transformation

$$P(x) = \ln\left[\frac{p(x)}{1 - p(x)}\right] = \alpha + \beta x$$

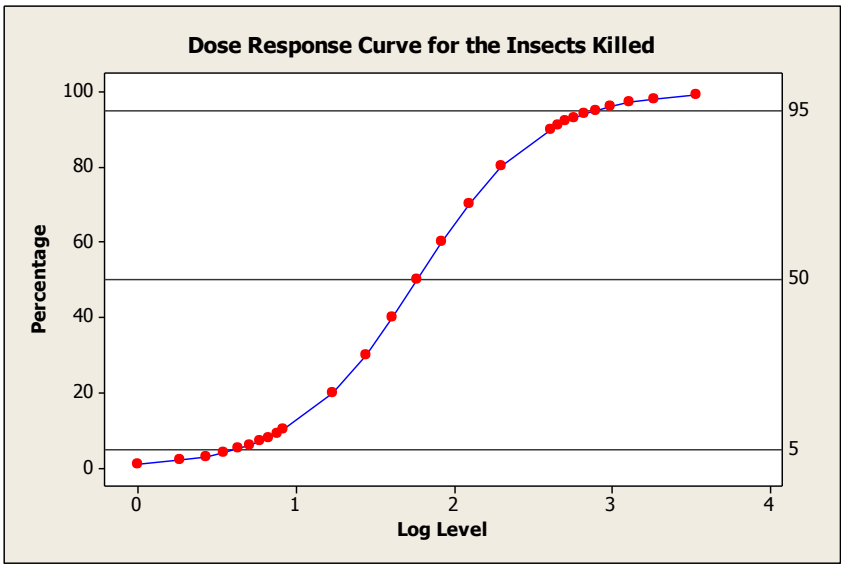
Probit Analysis: Insects Killed, Insects Treated versus Log Level

```
Distribution:   Logistic

Response Information

Variable         Value      Count
Insects Killed   Event         43
                 Non-event     57
Insects Treated  Total        100

Estimation Method: Maximum Likelihood
Regression Table
                      Standard
Variable       Coef     Error      Z      P
Constant   -4.60158   1.09142  -4.22  0.000
Log Level   2.60289  0.617895   4.21  0.000
Natural
Response          0

Log-Likelihood = -54.913

Goodness-of-Fit Tests

Method    Chi-Square  DF      P
Pearson      2.31729   6  0.888
Deviance     1.86373   6  0.932

Tolerance Distribution

Parameter Estimates
                     Standard    95.0% Normal CI
Parameter  Estimate      Error     Lower     Upper
Location    1.76787  0.0895252   1.59241   1.94334
Scale      0.384188  0.0912017  0.241256  0.611800
```

Table of Percentiles

Percent	Percentile	Standard Error	95.0% Fiducial CI Lower	Upper
1	0.0024829	0.418734	-1.51606	0.568583
2	0.272682	0.356317	-1.01417	0.756397
3	0.432398	0.319744	-0.718144	0.868057
4	0.546903	0.293733	-0.506334	0.948530
5	0.636655	0.273504	-0.340634	1.01193
6	0.710766	0.256931	-0.204076	1.06455
7	0.774098	0.242883	-0.0876118	1.10974
8	0.829553	0.230684	0.0141569	1.14953
9	0.879003	0.219900	0.104712	1.18519
10	0.923726	0.210235	0.186430	1.21764
20	1.23528	0.146607	0.747850	1.45147
30	1.44235	0.111599	1.10470	1.62320
40	1.61210	0.0930010	1.37227	1.78892
50	1.76787	0.0895252	1.58084	1.97797
60	1.92365	0.100575	1.74894	2.20751
70	2.09340	0.124571	1.90097	2.48877
80	2.30047	0.162807	2.06528	2.85303
90	2.61202	0.228350	2.29504	3.41853
91	2.65675	0.238163	2.32717	3.50056
92	2.70619	0.249091	2.36255	3.59141
93	2.76165	0.261429	2.40205	3.69346
94	2.82498	0.275613	2.44697	3.81020
95	2.89909	0.292319	2.49931	3.94703
96	2.98884	0.312682	2.56244	4.11300
97	3.10335	0.338829	2.64264	4.32509
98	3.26307	0.375544	2.75401	4.62140
99	3.53326	0.438124	2.94150	5.12361

Level of Insecticide	Insects Treated	Insects Killed	Log Level	Probability
1.6	10	1	0.47000	0.032983
3.0	15	2	1.09861	0.149057
4.0	12	3	1.38629	0.270279
5.0	15	5	1.60944	0.398339
6.0	13	6	1.79176	0.515538
7.0	8	5	1.94591	0.613823
8.0	17	13	2.07944	0.692318
9.6	10	8	2.26176	0.783391



The dose-response curve and the table of percentile are given. They show that for this data we obtain $LD_{50} = 1.76787$, $LD_5 = 0.63667$ and $LD_{95} = 2.89909$.

7. Bioassay

Bioassay, or biological assay, is a body of methodology concerned with assessing the potency of a stimulus (e.g., a drug, a hormone, or radiation) in its effect on (usually) biological organisms. Biological assays are methods for the estimation of nature, constitution, or potency of a material (or of a process) by means of the reaction that follows its application to living matter.

Qualitative Assays	Quantitative Assays
These do not present any statistical problems. We shall not consider them here.	These provide numerical assessment of some property of the material to be assayed, and pose statistical problems.

Definition: An assay is a form of biological experiment; but the interest lies in comparing the potencies of treatments on an agreed scale, instead of in comparing the magnitude of effects of different treatments.

This makes assay different from varietal trials with plants and feeding trials with animals, or clinical trials with human beings. The experimental technique may be the same, but the difference in purpose will affect the optimal design and the statistical analysis. Thus, an investigation into the effects of different samples of insulin on the blood sugar of rabbits is not necessarily a biological assay; it becomes one if the experimenter's interest lies not simply in the changes in blood sugar, but in their use for the estimation of the potencies of the samples on a scale of standard units of insulin. Again, a field trial of the responses of potatoes to various phospatic fertilizers would not generally be regarded as an assay; nevertheless, if the yields of potatoes are to be used in assessing the potency of a natural rock phosphate relative to a standard superphosphate, and perhaps even in estimating the

availability of phosphorus in the rock phosphate, the experiment is an assay within the terms of the description given herein.

History of biological assay

In the Bible, in the description of Noah's experiment from his ark by sending a dove repeatedly until it returns with an olive leaf, by which Noah knows or estimates the level of receding waters from the Earth's grounds, we find that it has all the three essential constituents of an assay – namely "stimulus" (depth of water), "subject" (the done) and "response" (plucking of an olive leaf).

Serious scientific history of biological assay began at the close of 19th century with Ehrlich's investigations into the standardization of diphtheria antitoxin. Since then, the standardization of materials by means of the reactions of living matter has become a common practice, not only in pharmacology, but in other branches of science also, such as plant pathology. However the assays were put on sound bases only since 1930's when some statisticians contributed with their refined methods to this area.

Structure of a Biological Assay

The typical bioassay involves a stimulus (for example, a vitamin, a drug, a fungicide), applied to a subject (for example, an animal, a piece of animal tissue, a plant, a bacterial culture). The intensity of the stimulus is varied by using the various "doses" by the experimenter. Application of stimulus is followed by a change in some measurable characteristic of the subject, the magnitude of the change being dependent upon the dose. A measurement of this characteristic, for example, a weight of the whole subject, or of some particular organ, an analytical value such as blood sugar content or bone ash percentage, or even a simple record of occurrence or non-occurrence of a certain muscular contraction, recovery from symptoms of a dietary deficiency, or death — is the response of the subject.

Types of Bioassays

Three main types (other than qualitative assays)are :

(i) DIRECT ASSAYS

(ii) INDIRECT ASSAYS

Direct Assays

We shall first take up DIRECT ASSAYS. In such assays doses of the standard and test preparations are sufficient to produce a specified response, and can be directly measured. The ratio between these doses estimates the potency of test preparation relative to the standard. If Z_S and Z_T are doses of standard and test preparations producing the same effect, then the relative potency ρ is given by

$$\rho = \frac{Z_S}{Z_T}$$

Thus, in such assays, the response must be clear-cut & easily recognized, and exact dose can be measured without time lag or any other difficulty.

A typical example of a direct assay is the "cat" method for the assay of digitalis. Preparation is infused until its heart stops (causing death). The dose is immediately measured. It can take two basic forms: estimating stimulus response; and evaluating the potency of one stimulus (e.g., a new drug or pollutant) relative to another (a 'standard' or familiar form).

Preparations	Tolerances	Mean
Strophanthus A (Test Prep.) (in .01 cc/kg.)	1.55, 1.58, 1.71, 1.44, 1.24, 1.89 2.34	1.68
Strophanthus B (Stan. Prep.) (in .01 cc/kg.)	2.42, 1.85, 2.00, 2.27, 1.70, 1.48, 2.20	1.99

Hence the estimated relative potency is $\hat{\rho}$ = 1.99/1.68 = 1.18

Thus 1 cc of tincture A is estimated to be equivalent to 1.18 cc of tincture B.

Indirect Assays

The estimation of stimulus response is often done using the generalized linear models. Let us now consider the problem of ***relative potency***.

Regression methods and maximum likelihood are used extensively in bioassay. Consider two stimuli, *T* and *S*. They are said to have similar potency if the effect on a response variable *Y* of some level *Z* of the stimulus is such that there is a constant scale factor ρ such that

$$Y_T(Z) = Y_S(\rho Z)$$

This implies that the stimulus-response relationship is essentially the same for both of the stimuli; we have only to scale the dose level by a simple multiplicative factor to obtain identical relationship.

In many cases it proves reasonable to assume that *Y* has a linear regression relationship with $x = \ln z$. Then we can write

$$E[Y_T(Z)] = \alpha + \beta x = \alpha + \beta \ln z$$

and if *S* is similar in potency to *T*, we have

$$E[Y_S(Z)] = \alpha + \beta \ln \rho z = \alpha + \beta \ln \rho + \beta \ln z$$

Thus we have two linear regression models of *Y* on ln *z* with the same slope but different intercepts. So we could seek to test if two stimuli have similar potency by carrying out a hypothesis of parallelism of the two regression lines; if this is accepted we can proceed to estimate ρ. This procedure is known as a parallel-line assay; it assumes that the regression lines are parallel and separated by a horizontal distance

$$\beta \ln \rho = \alpha_S - \alpha_T$$

Suppose we have random samples of size n_T and n_S of stimulus and response for the two stimuli, *T* and *S*. If the errors are uncorrelated and the error distribution is normal with constant variance then the maximum likelihood (or the least squares) estimators need to be chosen to minimize

$$R = \sum_{i=1}^{n_1} (y_{Ti} - \alpha - \beta\, x_{Ti})^2 + \sum_{i=1}^{n_2} (y_{Si} - \alpha - \beta\, x_{Si})^2$$

which yields

$$\hat{\beta} = \frac{\sum_{i=1}^{n_1+n_2} x_i y_i - n_T\, \bar{x}_T \bar{y}_T - n_S\, \bar{x}_S \bar{y}_S}{\sum_{i=1}^{n_1+n_2} x_i^{\,2} - n_T\, \bar{x}_T^{\,2} - n_S\, \bar{x}_S^{\,2}}$$

$$\hat{\alpha}_T = \bar{y}_T - \hat{\beta}\, \bar{x}_T\,,\ \hat{\alpha}_S = \bar{y}_S - \hat{\beta}\, \bar{x}_S \text{ and } \hat{\rho} = \exp\left(\frac{\hat{\alpha}_S - \hat{\alpha}_T}{\hat{\beta}}\right)$$

Example: The following table considers a bioassay problem concerning two treatment regimes applied to laboratory rats.

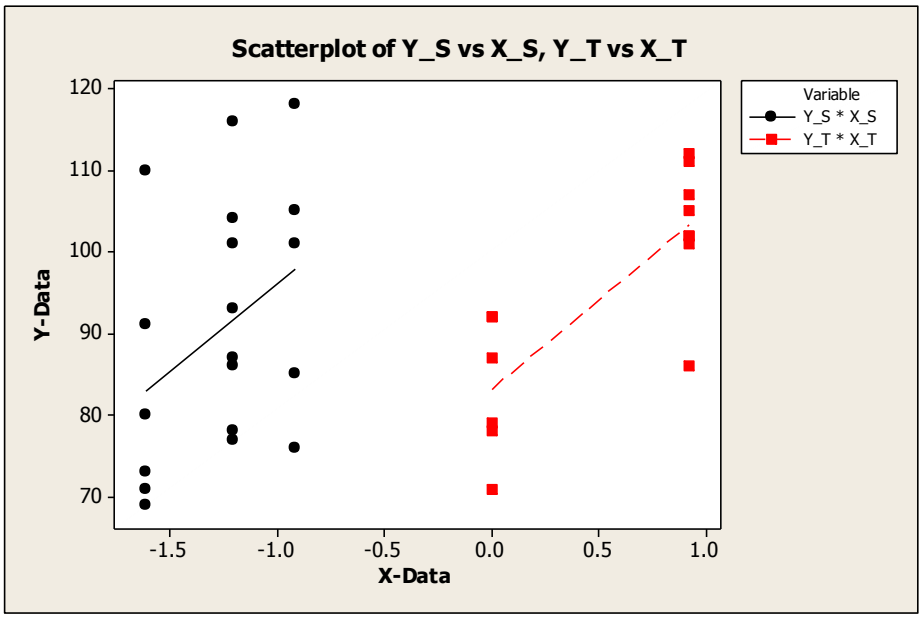
Table. Bioassay data on rats: uterine weights (coded)

Dose	Standard Treatment *S*			New Treatment *T*	
	0.2	**0.3**	**0.4**	**1.0**	**2.5**
	73	77	118	79	101
	69	93	85	87	86
	71	116	105	71	105
	91	78	76	78	111
	80	87	101	92	102
	110	86		92	107
		101			102
		104			112

For this data we obtain

$$\sum_{i=1}^{n_1+n_2} x_i y_i = \text{-1375.95},\quad \sum_{i=1}^{n_1+n_2} x_i^{\,2} = 38.0528,\ \bar{y}_T = 94.64,\ \bar{x}_T = 0.524,\ \bar{y}_S = 90.58,\ \bar{x}_S = \text{-1.2563}$$

$$n_T = 14,\quad n_S = 19,\ \hat{\beta} = 21.7673,\ \hat{\alpha}_T = 83.2340,\ \hat{\alpha}_S = 117.926,\quad \hat{\rho} = 4.92233$$



Repeated Measures

Example: The following table gives plasma floride concentration for litters of baby rats of different ages at different times after injectionof different doses of a drug.

Age (Days)	Dose (mg)	Post-injection time (min)		
		15	**30**	**60**
6	0.50	4.1	3.9	3.3
6	0.50	5.1	4.0	3.2
6	0.50	5.8	5.8	4.4
6	0.25	4.8	3.4	2.3
6	0.25	3.9	3.5	2.6
6	0.25	5.2	4.8	3.7
6	0.10	3.3	2.2	1.6
6	0.10	3.4	2.9	1.8
6	0.10	3.7	3.8	2.2
11	0.50	5.1	3.5	1.9
11	0.50	5.6	4.6	3.4
11	0.50	5.9	5.0	3.2
11	0.25	3.9	2.3	1.6
11	0.25	6.5	4.0	2.6
11	0.25	5.2	4.6	2.7
11	0.10	2.8	2.0	1.8
11	0.10	4.3	3.3	1.9
11	0.10	3.8	3.6	2.6

Solve the problem using the split-plot design