A Table of Contents

## Preliminaries

If $x_1, x_2, \ldots x_n$ are symbols, representing numbers, we may speak of this set of symbols as $x_i$, $i = 1,2,\ldots n$. If we wish to write a set of directions for adding these numbers, we may express them in the form $\sum_{i=1}^{n} x_i$, which is to be taken to mean $(x_1 + x_2 + \ldots + x_n)$. This use of the symbol $\sum$ is said to be summation notation.

Verify each of the following statements by writing it without the $\sum$ - notation.

1. $\sum_{i=1}^{n} c = n\,c$, $c$ any constant.

2. $\sum_{i=1}^{n} c\,x_i = c \sum_{i=1}^{n} x_i$

3. $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

4. $\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{m} y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i\,y_j$

Writing $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, verify the following statements using 1, 2, 3.

5. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

6. $$\sum_{i=1}^{n} (x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

7. $$\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i\,y_i - n\,\bar{x}\,\bar{y} = \sum_{i=1}^{n} x_i\,y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Formulae 6 and 7 are useful for making calculations when the $x_i$ and $y_i$ are given numerical values.

Further notation

1. The symbol $n!$ , $n$ any integer, to be read n factorial, means the product of all the integers up to and including $n$ , that is, $n! = n(n-1)(n-2)\ldots3.2.1$ .

   When $n = 0$ , $n!$ is to be taken to have the value 1 .

2. The symbol $\binom{n}{x}$ stands for the binomial coefficient

$$\frac{n!}{x!\,(n-x)!} = \frac{n(n-1)\ldots(n-x+1)}{x(x-1)\ldots3.2.1} \quad .$$

This number, necessarily an integer, is called the number of combinations of $n$ objects, $x$ at a time (also symbolized by $n\,C_x$ or $C_x^n$) , the number of possible selections of $x$ objects from $n$ , and the number of different arrangements, in a sequence, of $n$ objects, $x$ of which are indistinguishable from one another and the remaining $(n-x)$ of them are indistinguishable from one another.

3. $|x+y|$ stands for the absolute value of $x + y$ . For example, $|3-2| = 1$ , $|2-3| = 1$ .

Objectives

It may be convenient to speak of the purpose of statistics under two headings, even though they are by no means distinct:

(i) to reduce a collection of observations, perhaps large and unorganized, so that one can recognize answers to the questions that led to the making of the observations; to simplify the set of observations as much as is warranted without over-simplifying to the extent of suppressing or distorting information;

(ii) to make the observations in such a way as to support trustworthy and indisputable conclusions.

The study of (i) is a study of statistical techniques; (ii) is discussed under two headings, design of experiments and theory of sampling.

Obviously, (ii) takes precedence over (i), because a study of improperly taken observations can only lead to mistakes. It is convenient, though, to begin with (i) .

The population and the sample.

In all instances, a set of observations will be called a sample, the implication being that not all the observations that might have been made actually were made. The totality of observations that might, in principle, be made is called the population. It is about the population that we want to reach conclusions, but for reasons that are practical and insurmountable, we can get only a sample.

Often the population is genuinely infinite. In most other instances, it is large compared with the sample to be drawn from it and no violence is done in treating it as infinite, in the sense that it is not changed appreciably by the sampling.

The notion of population is largely conceptual and in practice the population is often difficult to define, in the sense of being able to pronounce that an item is or is not a member of the population we wish to study. For example, in an election poll, the population we would like to study is the population of people who will vote, but we are obliged to study the population of people eligible to vote, which may be substantially different.

As another example, the actuary makes a mortality table from the records of people who have died, but obviously he does so in

the belief that they represent adequately people who have not died and who, indeed, may not yet be born. Here he depends heavily on experience which says that these populations are stable, in the sense that they change only slowly with time. This dependence on stability is crucial and is not to be taken lightly. Only empirical evidence can demonstrate stability.

In these examples, the impression may have been left that populations are aggregates of people. Actually, the people are only bearers of numbers or other marks and these marks constitute the population. It is convenient, though, not to observe this distinction too rigidly.

The organization of samples.

For the moment, let us act as if our samples are large, so the question of organizing our numbers is real and important. It is worthwhile to keep in mind that our numbers may be reached in one or the other of two ways, by counting and by measurement.

Example 1.

Let us say that we have gathered up a sample of 100 ten-year-old boys, with a view to ascertaining from each a count of the

number of teeth missing. Each observation is reached by counting and is an integer , 0 , 1 , 2 etc. Our observations, then, consist of 100 integers, perhaps in a list, perhaps each written on a card.

To reduce these numbers to a more compact form, we can count the number of 0's, the number of 1's, etc. and enter these counts in a list.

| Number of teeth missing | Number of children |
|---|---|
| 0 | 26 |
| 1 | 35 |
| 2 | 21 |
| 3 | 10 |
| 4 | 5 |
| 5 | 2 |
| 6 | 1 |
| 7 | 0 |
| . | . |
| . | . |
| . | . |
| 32 | 0 |
| | 100 |

We may speak of this list in a more formal way by saying : let x represent the variable we are studying (i.e. number of teeth missing) and $f_x$ the number of times x is counted in the sample.

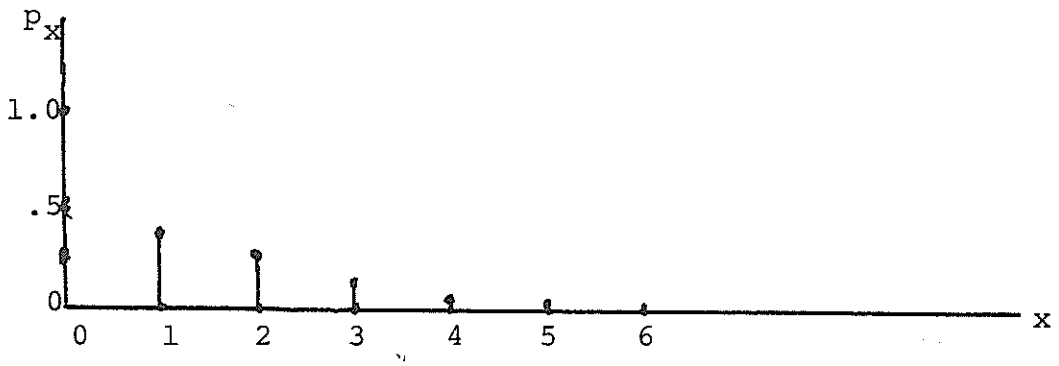
Then, the list $(x, f_x)$ , $x = 0, 1, 2, \ldots$ is the list enumerated above. x may be called a statistical variable and $f_x$ the frequency with which x occurs in the sample. The list is called a frequency distribution.

For many purposes, a more useful list is one that displays the proportions of the sample falling in the various categories, that is $\left(x, \frac{f_x}{n} = p_x\right)$, where n represents the sample size.

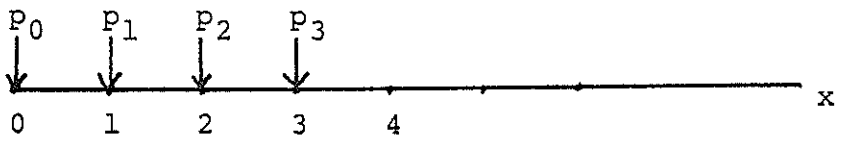
| x | $p_x$ |
|---|---|
| 0 | .26 |
| 1 | .35 |
| 2 | .21 |
| 3 | . |
| . | . |
| . | . |
| . | . |
| | 1 |

This form of the list may be called the relative frequency distribution or , if there is no chance of confusion, simply the frequency distribution. The size of the sample is not included in this list, and must be recorded separately.

A graphical representation of the frequency distribution can be made by plotting the points $(x, p_x)$ . These points should not be joined, but they may be emphasized by drawing vertical bars.



Another pictorial representation, not actually a graph, may be made by taking an axis of $x$ and putting a weight $p_x$ at each point $x$. This provides a mechanical analogue to the frequency distribution, which will provide some clues to further reduction of the frequency distribution.




In making up the frequency distribution, we have lost nothing. It represents simply a rearrangement of the raw records. If we make the reasonable supposition that the counts were made with strict accuracy, the frequency distribution gives a strictly accurate account of the sample. The sample, however, cannot be considered to give a strictly accurate representation of the population. The proportions $p_x$ in the sample are presumably different from the corresponding proportions in the population, owing to the vagaries of sampling. We will want to say that the sample proportions are wrong because of sampling error. Precautions which must be observed, in the taking of the sample, to make possible the discussion of

the difference between the population and the sample in terms of error will be gone into later on.

Example 2.

Let us speak again of a sample of 100 boys, but let us say now that each boy supplies a number obtained by measurement, say the measurement of his height. Again our sample is a set of 100 numbers, but this time they are not integers. Indeed, it is a convenient abstraction to think of height as a continuously varying quantity, even though we cannot measure it on a continuous scale.

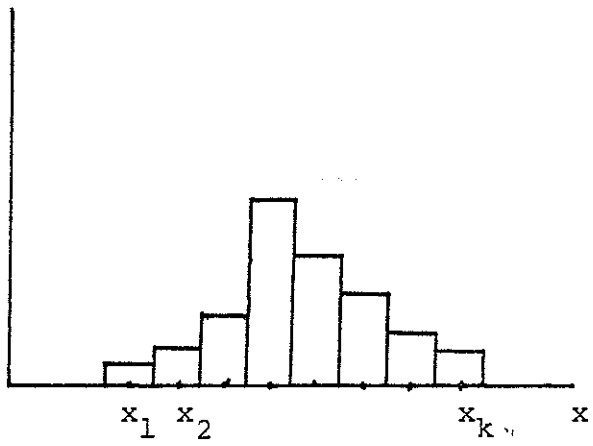
If we want to construct a frequency distribution from the sample, we must break up the range of heights into intervals, presumably equal in length and count the number of measurements falling in each interval. This is called grouping.

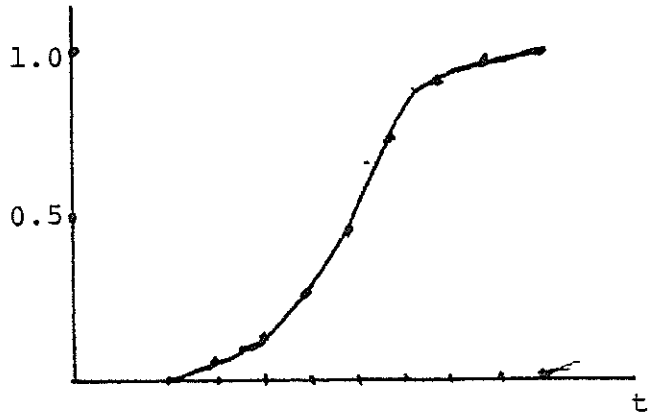
In contrast with counts, measurements are made with some inaccuracy and still more inaccuracy is introduced by grouping. Even so, these inaccuracies should be trivial compared with the variation in the heights themselves. Except in the physical sciences, inaccuracies in measurement are, or can be made to be, inconsequential compared with other sources of error.

A reasonable working rule is to group into 10 - 20 groups. Having grouped, it is customary to assign to each member of a group the value of the measurement equal to the mid-point of the interval. In choosing intervals, then, we should aim for simple mid-points and simple boundaries.

| interval | mid-point = x | frequency | relative frequency |
|---|---|---|---|
| - - - | $x_1$ | $f_{x_1}$ | $p_{x_1}$ |
| - - - | $x_2$ | $f_{x_2}$ | $p_{x_2}$ |
| - - - | $x_k$ | $f_{x_k}$ | $\frac{p_{x_k}}{1}$ |
| | | $N$ | |

Here again we can make graphical representations of the frequency distribution. The histogram is constructed by laying out the end-points of the intervals on an axis of $x$ and constructing rectangles on them whose areas are equal to the corresponding frequencies. The cumulative frequency distribution or ogive is constructed by plotting $\sum_{x \le t} p_x$ against $t$.



In constructing the ogive, it is reasonable to join the plotted points by a smooth curve.

The ogive is useful for answering questions like the following.

What value of a measurement is greater than or equal to exactly half of the measurements in the sample? This is the value of $t$ corresponding to an ordinate $\frac{1}{2}$. It is called the median of the sample.

On the difference between a population and a sample drawn from it.

1. Bias

Samples are sometimes taken in a way that persistently misrepresents the population that is ostensibly being sampled. This can come about in all sorts of ways, some of them subtle and unsuspected. An example of a dangerous instance of this is the questionnaire, mailed or handed out, leaving to the person who receives it the decision whether to answer and return it. Leaving aside the irresponsibility that can enter into the responses, ambiguity in the questions and the like, the population being sampled here is the population of people who choose to return the questionnaire, which may be grossly different from the population about which conclusions are sought.

A sampling procedure which persistently misrepresents the population is said to be biased.

2. Error.

Even in the absence of bias, a sample is sure to differ from the population, simply because no subset can be identical with the whole in all respects. Differences arising in this way may be big or small and there is no way of knowing the extent to which a particular sample differs from the population. We can, however, assess a sampling procedure and predict (in a sense yet to be described) how frequently differences of a given magnitude may be expected. In order to carry out this program, it is essential that the sampling be carried out in such a way that the theory of probability can be invoked to describe the behaviour of samples. For this reason, we will turn presently to the theory of Probability, which will provide simple models against which to compare real populations and clues to the conditions to be observed in the taking of samples. Before doing so, let us look at a further reduction of a sample, often made. In this, we follow a clue provided by the mechanical analogy to the frequency distribution.

In mechanics, a most useful notion has been that of the centre of gravity. This, in turn, depends on the notion of turning moment, which we shall call here first moment. If weights $p_x$ are

located at points $x$ along a line, the first moment of the system (about the origin) is $\sum x\, p_x$ .

The centre of gravity is that point at which the whole weight, if concentrated at the centre of gravity, would produce the same first moment. If $\bar{x}$ denotes this position, we have

$\bar{x} \sum p_x = \sum x\, p_x$ .

If we apply the same calculation to the frequency distribution, $\sum p_x = 1$ and $\bar{x} = \sum x\, p_x$ . $\bar{x}$ is called the average of the distribution. $\bar{x}$ is also the first moment about the origin. It is seen too that $\sum x\, p_x$ is simply the sum of the numbers in the sample (in the case of counts, at least) divided by the sample size, so that $\bar{x}$ , the average of the distribution, is simply the average of the sample, whether or not we choose to rearrange it into a frequency distribution.

Evidently the average tells us nothing about the distribution itself. It only tells us where the distribution is, i.e. its location. This is a drastic reduction indeed, if we think of replacing the distribution by its average and one may surmise that it is usually too drastic. So it is. At least, we need something, in addition, to express some important features of the distribution

itself. With this in mind, the following definition is made.

The $k^{th}$ moment of the distribution about a point $x_0$ is $\sum(x - x_0)^k p_x$. In particular, if $x_0$ is taken to be $\bar{x}$, the $k^{th}$ moment about $\bar{x}$ reflects some intrinsic feature of the distribution, inasmuch as its value does not depend on the location of the distribution. The most important of these numbers corresponds to $k = 2$, which yields the second moment about the average, $\sum(x - \bar{x})^2 p_x$. This number indicates the spread of the distribution, large values arising from widely-spread and diffuse distributions, small values indicating narrow and highly concentrated distributions. [Question: if the second moment about the average of a particular distribution has value zero, what kind of distribution is it?] This is, of course, an important feature of the distribution, but it must be recognized that the second moment about the average says nothing about the shape of the distribution. To speak of shape, moments of higher degree are needed. This will not be pursued here, but it is worth noting that if a distribution is symmetrical about its average, the third moment about the average has value zero. Note also that the first moment about the average, $\sum(x - \bar{x}) p_x$, necessarily has value zero.

The introduction of the notion of moments of a distribution has to do with the reduction of a sample to a few important numbers which summarize what the sample has to say about questions we want to ask about the population. This direction of thought will be taken up later, after the introduction of the notion of probability.

Probability

It is sometimes said that probability theory provides a way of dealing with uncertainty or with the unknown. There may be some grounds for such statements, but they are not useful. Obviously no theory can cope with every instance of uncertainty or lack of knowledge. Indeed, the first step is to recognize those instances in which the reasons for the uncertainty can be identified and made the basis for useful and verifiable predictions. This is most easily accomplished in gambling games, for reasons that will be mentioned later.

To make an example, think of a bag known to contain equal numbers of black and white beads. One bead is to be drawn from the bag. The outcome of this drawing is unknown, to be sure; indeed, it is unknowable. Is there anything we can say, then, about the outcome of the drawing, in view of our knowledge that blacks and whites are equally numerous? There is a strong intuitive appeal to the notion of defining the probability of getting a black bead to be $\frac{1}{2}$, i.e. the proportion of blacks in the bag, but it is important to recognize that there is no sensible answer to the question as it has been asked. This arises because of what we do not know. For

example, it may be that all the blacks are on top, or they may be distributed throughout the bag in any manner whatever and the way in which the drawing will be made will largely determine the outcome, regardless of the proportion of blacks in the bag. Clearly what is needed here is a procedure for drawing the bead that will not reflect the initial distribution of beads and therefore reflects only the proportion of black beads. A procedure which possesses this property will be said to be a random procedure.

In this instance, we know a way of meeting this requirement. Before making the drawing, we will mix the beads thoroughly. We might also say that the object of the mixing is to give each bead the same chance of being drawn as that of every other bead. With this proviso, then, we will define the probability of drawing a black bead to be $\frac{1}{2}$, i.e. the proportion (or relative frequency) of black beads. Similarly, if it is known that the bag contains twice as many blacks as whites, we define the probability of getting a black in one randomly made drawing, to be $\frac{2}{3}$.

The bag of beads here is a population. The drawing of a bead is the taking of a sample of one observation. We know enough about

the population to make a complete list of all the possible outcomes of our sampling, in this instance black, white. This list is called a sample space. (Often the items in the list are numbers and can be plotted in a geometrical space.) In any event, the items in the list are often called sample points.

We also know the frequencies in the population that determine the probability of each item (or point) in the sample space, provided the sampling is carried out randomly.

| sample point | probability |
|---|---|
| black | $\frac{2}{3}$ |
| white | $\frac{1}{3}$ |
| | 1 |

This list, describing randomly chosen samples of 1 , also describes the population, if we substitute the word frequency for probability.

We must, of course, think of more elaborate samples than samples of 1. If, for example, we propose to draw a sample of 2 beads, the sample space could be

| | | Number of blacks |
|---|---|---|
| black, black<br>black, white | or | 2 |
| white, black | | 1 |
| white, white | | 0 |

Thus, the notion of sample space does not dictate any one form. The list of possible outcomes can be made up in several different ways. We would naturally choose the one (or ones) most suitable to the questions we are asking.

The attaching of probabilities to sample points, assuming randomness in the sampling, requires computations using the calculus of probabilities, which will be introduced later.

To sum up the points of view put forward here, the notion of probability depends on two essential elements, a well-defined population with known frequencies and a procedure for drawing samples randomly. The notion of randomness is essentially undefined, although the objectives are clear enough and ways of achieving randomness must be devised in every application. In gambling games, both elements are well provided for, but in a wider context we can encounter real difficulties with each of them.

In one class of instances, like the bag of beads, in which we can envisage a finite number of items, each of which bears one or another of a set of marks (the marks constituting the population), randomness in the sampling takes the form of ensuring somehow that

all sets of items, of the size dictated by the sample, are equally probable. In these instances, there is a simple and natural definition of probability, which will be given shortly.

Events.

Think of any population of the sort singled out above, say the bag of beads and any random sampling, for example, the drawing of 5 beads. Any outcome, or any collection of outcomes, will be called an event.

Examples: 2 white and 3 black; at least one black; 5 black or 5 white. An event, then, is simply a set of sample points made up to correspond to some question that has been raised. It is convenient to extend the use of the word event to the two extremes, in which (1) the set of sample points includes the whole sample space; (2) the set of sample points contains no sample point. An example of (1), in the sampling of 5 beads from a population of black and white beads, is the event : at least one black or at least one white. An example of (2) is the event : 5 whites and 5 blacks. In (1), the event is certain to occur; in (2) the event cannot occur.

Capital letters will be used to symbolize events: A, B, E, X, etc.

A definition of probability

A random sampling can result in one and only one of a set of $N$ equally probable outcomes. Of these, $n$ produce an event $A$ and the other $N - n$ possible outcomes do not produce $A$. The probability that $A$ will occur is defined to be $\frac{n}{N}$. The probability will be given the symbol $P(A)$. Thus, $P(A) = \frac{n}{N}$.

With populations of the sort for which this definition has meaning, we can calculate the probability of any event whatever simply by counting to obtain $n$ and $N$. The simplest probability calculations are carried out so, but when the events are complicated, direct counting can be difficult and confusing and we can get real help from a few rules which are easily derived from the definition. These rules reduce the calculation of the probability of a complicated event to a combination of the probabilities of simpler events.

Some examples.

The population: a deck of playing cards.

The sample: a sample of 1 card to be drawn randomly.

(1) The event: a heart or a spade will be drawn. By a direct count, $P(\text{heart or spade}) = \frac{26}{52} = \frac{13}{52} + \frac{13}{52} = P(\text{heart}) + P(\text{spade})$. If we let $A$ represent the event: a heart will be drawn and $B$: a spade

will be drawn, we can write the solution as

$$P(A \text{ or } B) = P(A) + P(B) .$$

We might say that we have "decomposed" the event A or B into a combination of A and B and our rule combines the probabilities of A and of B to give the probability of A or B .

(2) The event: a heart or a face card will be drawn.

By a direct count, there are 22 out of the 52 cards that produce this event. We might also carry out the count as follows: there are 13 hearts and 12 face cards; the probability of getting a heart (A) is therefore $P(A) = \frac{13}{52}$ ; and of getting a face card (B) is $P(B) = \frac{12}{52}$ . The required probability is not, however, $P(A) + P(B)$ , as it was in example (1) , because we have counted the face cards that are hearts both as face cards and as hearts. Clearly, then we have

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) .$$

This seems to be not a long step forward. We have expressed the probability of one complicated event, A or B, in terms of the probability of another complicated one, A and B. It is a fact, though, that often one of the complicated ones is easily calculated and the rule becomes most useful.

The distinction between example (1) and example (2) is

important. The two simple events in (1) are such that if one of them happens the other cannot. Events of this sort are said to be mutually exclusive. In example (2), the two simple events can both happen.

The use of the words and and or in these examples deserves some comment. In example (2),

A or B means at least one of A and B and

A and B means both A and B. In example (1),

A or B implies exactly one of A and B, because no other possibility exists, while A and B is impossible.

In example (2), we could ask for the probability of exactly one of A and B . This event is composed of two simpler ones, (A and not B) and (B and not A). The event (exactly one of A and B) is the same as (A and not B) or (B and not A).

$$P(\text{A and not B}) = \frac{10}{52}, \qquad P(\text{B and not A}) = \frac{9}{52},$$

and these two events are mutually exclusive. Therefore,

$$P(\text{exactly one of A and B}) = \frac{10}{52} + \frac{9}{52}.$$

Notation.

Arguments of the sort we have been going through are greatly facilitated by the use of a good notation to take the place of the

most important words that have been needed to describe events and make statements about them.

1. not.

When we have an event A and wish to specify that A must not occur, we will write this as $\bar{A}$ (sometimes called the complementary event to A). Clearly A and $\bar{A}$ are mutually exclusive and one of these must happen, that is, A and $\bar{A}$ are exhaustive (the events A and $\bar{A}$ exhaust the possibilities). Since "certainty" obviously has probability 1, we have

$$P(A \text{ or } \bar{A}) = P(A) + P(\bar{A}) = 1 \quad . \quad P(\bar{A}) = 1 - P(A).$$

2. or.

If we want to replace or by a symbol, one common notation is to use + , with the clear understanding that the symbol used so does not imply arithmetical addition. On the other hand, one of the rules for P(A or B) becomes $P(A + B) = P(A) + P(B)$ , a kind of "addition" rule. The + on the left means or and the + on the right means arithmetical addition.

3. and.

The symbol to be used as a replacement for and will be a multiplication symbol. A and B then becomes A,B or simply AB . One of the rules for P(A or B) now becomes

$$P(A+B) = P(A) + P(B) - P(AB) \quad .$$

4. certainty and impossibility

It is convenient to have symbols for these special "events".

The symbols 1 and 0 will be used here.

Then, $P(1) = 1$ , $P(0) = 0$ , obviously.

5. is the same as the event

We had occasion earlier to say that two different descriptions specified the same event. We shall use the symbol = for this purpose.

We are now in a position to write statements about events wholly in symbols. For example,

$A + \bar{A} = 1$ means A and $\bar{A}$ are exhaustive.

$AB = 0$ means: A and B are mutually exclusive.

Further examples.

We have a rule for evaluating $P(A + B)$ which may involve the calculation of $P(AB)$ . This example is concerned with evaluating $P(AB)$ .

The population : a deck of playing cards.

The sample : one card to be drawn randomly.

Two events : A the card will be a heart.

B the card will be a face card.

The question : What is the probability that both A and B will happen?

We are asked for $P(AB)$ , which is easily seen by a direct count to be $\frac{3}{52}$ . The important question, though, is whether this answer can be given by some combination of $P(A)$ and $P(B)$ . The answer is : in general, it cannot.

We can calculate $P(A) = \frac{13}{52}$ . To convert this to the value of $P(AB)$ requires a multiplication by $\frac{3}{13}$ , i.e.

$$P(AB) = P(A) . \frac{3}{13} ,$$

Can we interpret $\frac{3}{13}$ as a probability in any useful way? The fraction $\frac{3}{13}$ is seen to be the probability of drawing a face card, provided the sampling is confined to the 13 hearts. It will be called the probability of $B$ conditional upon $A$ and given the symbol $P(B/A)$ . We have, then, $P(AB) = P(A)\ P(B/A)$ . In the same way, we could reach $P(AB) = P(B)\ P(A/B)$ . This rule will be called the multiplication rule. It states that the probability of a symbolic product, $AB$ , is the arithmetical product of the probabilities $P(A)$ , $P(B/A)$ .

Independence.

A most important special situation is that in which the conditional probability $P(B/A)$ is equal to the unconditional probability $P(B)$ . In words, the occurrence of $A$ does not affect

the probability of $B$ . When this relation holds, it is said that $B$ is independent of $A$ . If $B$ is independent of $A$, it follows that $A$ is independent of $B$ since $P(AB) = P(B)\ P(A/B) = P(A)\ P(B/A)$

$= P(A)\ P(B)$ when $P(B/A) = P(B)$ ,

we have $P(A/B) = P(A)$ , which says that $A$ is independent of $B$ . We may therefore simply speak of $A$ and $B$ as independent of each other and the multiplication rule becomes $P(AB) = P(A)\ P(B)$ .

In the case of independence, then, $P(AB)$ can be expressed entirely in terms of $P(A)$ and $P(B)$ .

Recapitulation.

We have two general rules, each with an important special case.

1. $P(A + B) = P(A) + P(B) - P(AB)$

1'. $= P(A) + P(B)$ when $A$ and $B$ are mutually exclusive.

2. $P(AB) = P(A)\ P(B/A) = P(B)\ P(A/B)$

2'. $= P(A)\ P(B)$ when $A$ and $B$ are independent.

These rules can be extended to deal with more than two events.

1'. $P(A + B + C) = P(A) + P(B) + P(C)$ when $A, B, C$ are mutually exclusive.

2. $P(ABC) = P(A)\ P(B/A)\ P(C/AB)$

2'. $= P(A)\ P(B)\ P(C)$ when $A, B, C$ are independent.

Rule 1 becomes rather complicated.

A few problems will illustrate ways in which the rules may be put to use.

Example.

Think of a game in which we have two boxes, the first containing 5 white and 15 black beads, the second 10 white and 10 black. The game is to be played by choosing one of the boxes by some probabilistic rule and then drawing randomly one bead from the box so chosen. The player wins if he draws a white bead.

Let us say that the box is to be chosen by throwing a die If it shows 1 or 2, we choose box 1, otherwise we choose box 2.

The most effective way of bringing the rules into play is to attach symbols to the various events that are encountered in the game.

Let $A_1$ represent the event : box 1 will be chosen.

Let $A_2$ represent the event : box 2 will be chosen.

(Note that $A_1 + A_2 = 1$ .)

Then, $P(A_1) = \frac{1}{3}$ , $P(A_2) = \frac{2}{3}$ .

Let $W$ represent the event : a white ball will be drawn. We seek the value of $P(W)$ .

By direct argument, we see that $W$ can be decomposed into two mutually exclusive and exhaustive events, $WA_1$, and $WA_2$. We can write $W = WA_1 + WA_2$ and $P(W) = P(WA_1 + WA_2) = P(WA_1) + P(WA_2)$

$$= P(A_1)\ P(W/A_1) + P(A_2)\ P(W/A_2)$$

Now $P(W/A_1) = \frac{1}{4}$, $P(W/A_2) = \frac{1}{2}$.

Therefore, $P(W) = \frac{1}{3} \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{2} = \frac{5}{12}$.

The essential feature of this solution is a particular instance of a general rule. In a given sampling, an event $W$ may happen. If it does, it must happen in conjunction with one and only one of a set of events $A_1, A_2 \ldots A_k$. (i.e. $A_1, A_2 \ldots A_k$ are mutually exclusive and exhaustive. In symbols, $A_1 + A_2 + \ldots + A_k = 1$, $A_i\ A_j = 0$ all $i \neq j$.) Then, $W = WA_1 + WA_2 \ldots + WA_k$ and $P(W) = P(WA_1) + P(WA_2) + \ldots + P(WA_k)$. This fact is sometimes given as a theorem.

Example.

In the foregoing example, another question can be answered, one which sounds bizarre, partly because of the way it is usually asked. If the game has been played and a white ball has been drawn, what is the probability that box 1 was chosen in the course of the game? This sounds like asking for the probability of a "cause".

In any event, it seems that we are asking for the value of $P(A_1/W)$ . Now, the multiplication rule states that

$$P(A_1W) = P(A_1)\ P(W/A_1) = P(W)\ P(A_1/W) .$$

Thus, $P(A_1/W) = \dfrac{P(A_1W)}{P(W)} = \dfrac{1/12}{5/12} = 1/5$

This use of the multiplication rule is sometimes called Bayes' Theorem.

Random Variables

In a random sampling of any sort from a population, we have envisaged a complete list of possible outcomes (the sample space) and the probabilities associated with them (which may require some calculations to obtain). This coupling of outcomes and probabilities is called a probability distribution. Often the outcomes are specified by numbers (e.g. the number of heads in 10 throws of a coin,) but whether this is so or not, there are many natural ways in which numbers become associated with sample points, for example, a payoff in a gambling game. We then have a list of numbers and the probabilities associated with them. If $x_i$ is one of the numbers in the list and $f(x_i)$ the corresponding probability, the set of numbers $x_i$, $f(x_i)$ specifies the probability distribution of a variable $x$. This variable, $x$, is called a random variable, because its values are defined on a sample space.

| $x$ | $f(x)$ |
|---|---|
| $x_1$ | $f(x_1)$ |
| $x_2$ | $f(x_2)$ |
| $x_i$ | $f(x_i)$ |
| $x_k$ | $f(x_k)$ |
| | 1 |

This list is seen to have the same form as the frequency distribution discussed earlier to describe a sample. Indeed, it describes a population in exactly the same sense. In statistical situations, the values of the $f(x_i)$ would not be known and the object of the sampling would be to estimate the values of the $f(x_i)$ or something equivalent. In probabilistic situations, the $f(x_i)$ are known or

calculable and can be used to make probabilistic predictions.

Clearly, we can make calculations with a probability distribution in the same way as we did with the frequency distribution of a sample, to get numbers descriptive of it. For example, the "centre of gravity" calculation yields $\sum_{i=1}^{k} x_i f(x_i)$ , which will be called the mean of the distribution, symbolized by $\mu$ , and the second moment about the mean, $\sum_{i=1}^{k} (x_i-\mu)^2 f(x_i)$ , called the variance of the distribution, with symbol $\sigma^2$ . The square root of the variance, $\sigma$, is called the standard deviation of the distribution.

Expectation

Calculations of this sort are nicely brought together by an operation called the calculation of the expectation (or mathematical expectation) of a random variable. If $x$ is any random variable, taking values $x_1, x_2, \ldots x_k$ with probabilities $f(x_1)$ , $f(x_2)$ , $\ldots f(x_k)$, the expectation of $x$ is defined to be

$$Ex = \sum_{i=1}^{k} x_i f(x_i) ,$$

that is, it is simply the mean of $x$ . With this notation, then, $\mu = Ex$ , $\sigma^2 = E(x-\mu)^2 = E(x-Ex)^2$ . Note that the expectation of a random variable is simply a number, descriptive of the distribution.

The notion of expectation is important because the operation has two important properties.

1. The expectation of a sum of random variables is equal to the sum of their expectations.

2. The expectation of a product of random variables is equal to the product of their expectations, provided they are independent.

Let $x$ and $y$ be two random variables, taking values $x_i$, $i = 1,\ldots k$ and $y_j$, $j = 1,\ldots m$ with probabilities $f(x_i)$ and $g(y_j)$. Then $x$ and $y$ are said to be independent if $P(x=x_i \text{ and } y=y_j) = f(x_i)\, g(y_j)$.

Proof of 1. when $x$ and $y$ are independent.

$x + y$ is a random variable taking $km$ values $x_i + y_j$, $i = 1\ldots k$, $j = 1\ldots m$ (not necessarily all different).

By definition, $E(x+y) = \sum_{i=1}^{k} \sum_{j=1}^{m} (x_i + y_j)\, f(x_i)\, g(y_j)$

$$= \sum_i \sum_j x_i\, f(x_i)\, g(y_j) + \sum_i \sum_j y_j\, f(x_i)\, g(y_j)$$

$$= \Big(\sum_i x_i\, f(x_i)\Big)\Big(\sum_j g(y_j)\Big) + \Big(\sum_j y_j\, g(y_j)\Big)\Big(\sum_i f(x_i)\Big)$$

$= Ex + Ey$, since $\sum_i f(x_i) = \sum_j g(y_j) = 1$.

The theorem may be proved without the condition of independence, in a somewhat more elaborate calculation.

Proof of 2.

$$Exy = \sum_i \sum_j x_i\, y_j\, f(x_i)\, g(y_j) = \Big(\sum_i x_i\, f(x_i)\Big)\Big(\sum_j y_j\, g(y_j)\Big)$$

$$= Ex\ Ey$$

These theorems are easily extended to sums and products of three or more random variables.

Verify from the definition of expectation that

1. $Ec = c$ , $c$ any constant.
2. $Ecx = c\ Ex$ .
3. $E(x - Ex) = 0$ .
4. $E(x - Ex)^2 = Ex^2 - (Ex)^2$ .
5. $E(ax + b) = a\ Ex + b$ .
6. $\mathrm{Var}(ax + b) = a^2\ \mathrm{Var}\ x$ .

5 and 6 show the effect of changing the origin and the unit of measurement of a statistical variable.

We will shortly be obliged to put to use the notion of the distribution of sums of random variables and in particular to be able to write quickly expressions for the means and variances of these distributions. The means we can already manage.

The mean of a sum is the sum of the means.

It will be convenient to have a similar rule for variances.

$$\text{Var}\ (x+y) = E\{x+y - E(x+y)\}^2 = E\{(x-Ex) + (y-Ey)\}^2$$

$$= E\{(x-Ex)^2 + (y-Ey)^2 + 2(x-Ex)(y-Ey)\}$$

$$= E(x-Ex)^2 + E(y-Ey)^2 + 2E(x-Ex)(y-Ey)$$

The term $E(x-Ex)(y-Ey)$ has not arisen before. It will be called the <u>covariance</u> of $x$ and $y$.

We have, then,

$$\text{Var}(x+y) = \text{Var}\ x + \text{Var}\ y + 2\ \text{Cov}\ (x,y)\ .$$

Now, if $x$ and $y$ are independent,

$$\text{Cov}\ (x,y) = E(x-Ex)(y-Ey) = E(x-Ex)E(y-Ey) = 0$$

Thus, when $x$ and $y$ are independent, the variance of their sum is the sum of their variances.

<u>Exercises</u>.

Verify, by making the appropriate calculations, the following rules. $x$ and $y$ represent random variables, $a$ and $b$ stand for any constants.

1. $E(ax + by) = aEx + bEy$ .

2. $\text{Var}(ax + by) = a^2\ \text{Var}\ x + b^2\ \text{Var}\ y + 2ab\ \text{Cov}\ (x,y)$.

3. $E(x-y) = Ex - Ey$ .

4. $\text{Var}(x - y) = \text{Var}\ x + \text{Var}\ y$ , if $x$, $y$ are independent.

The Binomial Population

A population in which each item bears one or the other of two marks (black and white, head and tail, pass and fail, 0 and 1) is called a binomial population. In all instances we can identify the marks with 0 and 1. A sample from a binomial population is then a set of 0's and 1's.

A binomial population is wholly specified by the proportion of 1's in it, if it is infinite. If it is finite and this fact must be taken into account, the total number of items in it must also be known.

Samples from an infinite binomial population

The population will be described by a variable $x$, which takes values 0 and 1, and the proportion of 1's and 0's, say $p$ and $1 - p = q$, in the population. In a statistical problem, $p$ is not known.

| $x$ | $f(x)$ |
|---|---|
| 1 | $p$ |
| 0 | $q$ |
| | 1 |

After a randomly chosen sample of size $n$ has been selected, we have a set of numbers which will be designated $x_1, x_2, \ldots x_n$. In this instance, each of the $x_i$ is 0 or 1. We may ask, then,

what can be learned or inferred from the sample about the population. The answer must be : very little, unless we view this sample in the context of all the samples of size $n$ that might eventuate in a random sampling of this population. To this end, we will regard the symbols $x_1, x_2, \ldots x_n$ not only as standing for numbers which we have obtained in an actual sample, but also as random variables describing what we can get in a sampling of this kind. In this instance, there are $2^n$ possible samples, each in some sense different from the others. The set of symbols $x_1, x_2, \ldots x_n$ represents all these $2^n$ samples.

The sample, when we get it, can be rearranged in the form of a frequency distribution, simply by counting the number of 1's and 0's .

| $\underline{x}$ | $\underline{f_x}$ | $\underline{p_x}$ |
|---|---|---|
| 1 | $n_1$ | $p_1$ |
| 0 | $\frac{n_o}{n}$ | $\frac{p_o}{1}$ |

Intuitively, one would expect $p_1$ to be "close to" the unknown proportion $p$ . Note that $n_1$ is simply the sum $\sum_{i=1}^{n} x_i$ and $p_1$ is $\frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$ . Hence, we are led to study the sum and average of the set of independent random variables $x_1, x_2, \ldots x_n$ . Let us write $y = \sum x_i$ and $\bar{x} = \frac{y}{n}$ .

$y$ is a random variable, defined as the sum of $n$ identical

independent, random variables $x_1, x_2, \ldots x_n$ . We do not know the distribution of $y$ (it will be derived later), but we do know the distribution of each of the x's. It is dictated by the population. We can write, for each of the $x_i$ :

$$E\, x_i = 1.p + 0.q = p$$

$$\text{Var } x_i = (1 - p)^2 p + (0 - p)^2 q = pq \ .$$

Therefore, using our rules:

$$Ey = np \qquad \text{and} \qquad E\bar{x} = p$$

$$\text{Var } y = npq \qquad \text{Var } \bar{x} = \frac{pq}{n}$$

We see, then, that $\bar{x}$ is a random variable whose mean is $p$ , the same as the mean of the population and whose variance is $\frac{pq}{n}$ , $\frac{1}{n}$ $\frac{\text{th}}{}$ the variance of the population. The distribution of $\bar{x}$ is thus more "closely packed" about the mean than is the population.

After the sample is taken and the average calculated, we can regard its value as a single observation drawn randomly from the "population" described by $\frac{y}{n} = \bar{x}$ . The "observed" value of $\bar{x}$ , then, ought to be closer to the mean (p) than an observation drawn randomly from the population.

The binomial distribution.

The random variable $y$ is the number of 1's in the sample. It may take each of the values $0, 1, \ldots n$. For any such value of $y$, the probability of getting $y$ 1's and $n - y$ 0's, in a given order, is $p^y q^{n-y}$ (independent). There are $\frac{n!}{y!(n-y)!} = \binom{n}{y}$ different events of this sort.

The probability of $y$ 1's is therefore

$$\frac{n!}{y!(n-y)!} \; p^y q^{n-y} \quad , \text{ (mutually exclusive)}$$

$y = 0, 1, \ldots n$.

This is the binomial distribution, so called presumably because each of these probabilities is a term in the binomial expansion of $(q + p)^n$.

The negative binomial distribution

The infinite binomial population may be sampled in another way. If items are drawn randomly, one at a time, until a 1 is drawn, the number of draws is a random variable, say $x$, whose possible values are $1, 2, 3, \ldots$. The probability of making $x$ draws is evidently $q^{x-1} p$. More generally, if the sampling is to terminate when the $k^{th}$ 1 appears, the probability of making $x$ draws is $\frac{(x-1)!}{(k-1)!(x-k)!} q^{x-k} p^k$, $x = k, k+1, \ldots$.

This kind of sampling is sometimes called inverse sampling.

It has some special uses, but will not be studied further here.

Sampling from a finite binomial population

If the population contains $N$ items, of which $Np$ are 1's and $Nq$ are 0's, the probability that a randomly chosen sample of $n$ items will contain exactly $y$ 1's is

$$\frac{\binom{Np}{y}\binom{Nq}{n-y}}{\binom{N}{n}} \quad , \quad y = 0, 1, 2, \ldots n \ .$$

This is usually called the hypergeometric distribution. The mean of $y$ is $np$ and the variance of $y$ is $npq\,\frac{N-n}{N-1}$. The average of the sample, $\frac{y}{n}$, has mean $p$ and variance $\frac{pq}{n}\,\frac{N-n}{N-1}$.

Some uses of the binomial distribution

Standard uses of the binomial distribution will be encountered from time to time later on. Two rather special uses will be mentioned here.

Acceptance sampling.

Purchasers often buy items (ball bearings, say) in large consignments. Inevitably, every consignment contains some fraction of defective items and the purchaser wants some protection against accepting consignments with too large a fraction defective. Usually, for one or another of several reasons (consignment too large,

testing expensive, testing destructive), complete inspection is not feasible and the decision must be based on a sample, each item of which is to be tested and declared either acceptable (0) or defective (1) . The population (consignment) is then binomial with an unknown fraction p of 1's .

The number of rules we may think of for rejecting consignments (because p is thought to be too large) is unlimited. The choice among them will be based on many things (cost of sampling, cost of testing and so on), but some of them are probabilistic.

To make an example, suppose we propose to accept or reject large consignments of life rafts by choosing randomly a sample of 10 rafts and testing them. If no defective raft is found in the sample, the consignment will be accepted; otherwise, it will be rejected. What sort of protection does this rule provide?

Represent by p the unknown proportion of defective rafts in the consignment. Then, the probability of accepting the consignment is $(1-p)^{10}$ . Call this $P_{10}$ . The graph of $P_{10}$ against p is called the operating characteristic curve of the plan. From it, we can read the probability of accepting consignments having any value whatever of the fraction defective.

These curves are useful in comparing plans with a view to deciding which to adopt.

Clearly, in any acceptance plan, we may make mistakes of two kinds.

I. We may reject "good" consignments.

II. We may accept "bad" consignments.

If we are to make these notions quantitative, we must be prepared to say what good and bad mean in this context. This is a wholly practical decision.

Let us say that we wish to accept consignments with $p \leq .01$ (good) and to reject consignments with $p \geq .05$ (bad). For the plan we have been discussing, we can calculate, or read from the O. C. curve, the probabilities of the two kinds of mistakes.

The probability of accepting consignments with $p = .01$ is .904 . Hence, the probability of making a mistake of type I is $1 - .904 = .096 = \alpha$ (say) . Similarly, the probability of accepting consignments with $p = .05$ is $.599 = \beta$ (say). This is the probability of making a mistake of type II. These two probabilities were determined when the acceptance plan was chosen. If we do not like them, we must devise another plan with more

suitable values of $\alpha$ and $\beta$. We would like them both to be small, but considerations of cost usually lead to compromises.

Estimation of the size of a population

In many situations, the possibility of conducting a census does not exist and in others, even though, in principle, possible, there may be strong reasons for not doing so. An instance is the determination of the number of fish in a lake. A device that has been used is to capture a number of fish, mark and release them and subsequently sample the population again. The fraction of the sample bearing marks can be used to estimate the size of the population.

If the size of the population is $N$ (unknown) and if $X$ (known) marked individuals have been introduced, marked and unmarked individuals make up a binomial population with parameter $p = \frac{X}{N}$ (unknown). If the sample of $n$ individuals proves to have $x$ marked, then $\frac{x}{n}$ estimates $p = \frac{X}{N}$. The equation $\frac{x}{n} = \frac{X}{N}$ yields an estimate of $N$.

The chief weakness of this procedure lies in our inability to sample randomly and instances of fantastically wrong estimates are abundant.

The procedure is called a mark-recapture method. There are many variants of it.

Computations with the binomial distributions

1. When $n$ is not large and $p$ is a simple number not too close to 0 or 1, the calculations are not difficult and are facilitated by available tables.

2. When $p$ is very small, then inevitably $n$ must be large. On both counts, the calculation of values of the binomial probabilities is next to impossible. In these circumstances, it is reasonable to seek an approximation when $n$ gets large and $p$ gets small in such a way that $np = m$, a constant. It is found that, under these conditions, $\binom{n}{y} p^y q^{n-y}$ approaches $e^{-m} \frac{m^y}{y!}$, a much simpler calculation. This is known as the Poisson distribution. It is extensively tabulated and is useful in dealing with rare events.

3. When $n$ is large and $p$ is not small, another kind of approximation is needed. To set this up, we must now turn attention to continuous random variables.

Continuous distributions

When observations are made by measuring, it is a convenient idealization to regard all values, within some range, as possible. Similarly, thinking of a population of men, we can think of their heights taking every value.

To see how to describe the distribution of frequencies in such populations, one can think of taking a sample, form a grouping and plot the histogram. Then think of adding to the sample indefinitely and at the same time diminish the grouping interval. In a limiting sort of way, the histogram becomes a smooth curve which, from the manner in which it was reached, has the property that the area under it between any two values of the statistical variable furnishes the proportion of the population with values in this range. The total area under the curve must be unity. The curve is called the density curve of the distribution.

To describe the curve in symbols, we may call the statistical variable $x$ and the density $f(x)$. The graph of $f(x)$ against $x$ is the density curve.

The notion of expectation can be extended to apply to continuous random variables. This will not be done here. It will be sufficient to accept the fact that its properties remain the same as in the discontinuous case. In particular, the theorems about expectations of sums and products remain true.

The central limit theorem

Continuous random variables may have any distribution whatever,

but one distribution, above all others, plays a central role in statistics, for both theoretical and empirical reasons. It is called the normal (sometimes Gaussian) distribution Its chief importance derives from a mathematical proof that averages (or sums) of samples from any population, continuous or discontinuous, have distributions that tend toward the normal form in reasonably large samples. The proof that this is so constitutes the central limit theorem. The proof requires the assumption that the first few moments of the population be finite, but this imposes no restriction to the use of the theorem with actual populations.

The normal distribution

For any random variable $x$, we define the mean, $Ex$, symbolized by $\mu$, and the variance, $E(x - \mu)^2$, symbolized by $\sigma^2$. The normal distribution is completely specified by its mean and its variance. The notation $N(\mu,\sigma^2)$ will be used to stand for "a normal distribution with mean $\mu$ and variance $\sigma^2$".

The normal curve is symmetrical about its mean, where it reaches its highest point. It is sometimes described as "bell-shaped". In addition to the central limit theorem and the fact that it is completely determined by its mean and variance, the normal distribution

has the further property that a sum of independent normal variables is normal.

Standardized random variables.

Corresponding to any random variable $x$, with mean $\mu$ and variance $\sigma^2$, we can find another of the same kind ($z$) having mean 0 and variance 1, by writing $z = \frac{x - \mu}{\sigma}$. $z$ is called a standardized random variable. This device is particularly useful for finding areas under any normal curve. If the area under a normal curve $N(\mu, \sigma^2)$ between $x_1$ and $x_2$ is required, we can move to the standardized curve $N(0,1)$ by writing $z = \frac{x - \mu}{\sigma}$, and find the area under the standardized curve between $z_1 = \frac{x_1 - \mu}{\sigma}$ and $z_2 = \frac{x_2 - \mu}{\sigma}$. Areas under the standardized normal curve have been extensively tabulated.

The normal approximation to the binomial.

If $y$ has a binomial distribution with parameters $n$ and $p$, where $n$ is large and $p$ is not too small, the central limit theorem asserts that $z = \frac{y - np}{\sqrt{npq}}$ is approximately $N(0,1)$. Therefore, if we seek $P(y_1 \leq y \leq y_2)$, the value of this probability is approximately the area under $N(0,1)$ between $z_1 = \frac{y_1 - np}{\sqrt{npq}}$ and $z_2 = \frac{y_2 - np}{\sqrt{npq}}$. The approximation is improved by finding the

area under $N(0,1)$ between

$$z_1 = \frac{y_1 - np - \frac{1}{2}}{\sqrt{npq}} \quad \text{and} \quad z_2 = \frac{y_2 - np + \frac{1}{2}}{\sqrt{npq}}$$

Tests of Significance

An example.

A poll of 20 voters yields 15 intentions to vote for candidate A and 5 intentions to vote against him. Does this provide strong evidence that he will get more than half of the votes?

We can think of the population of voters as a binomial population with an unknown proportion $p$ of voters who will vote for A . We must suppose that the sample of 20 voters came randomly from this population; otherwise it has no use for any purpose. The question to be settled, then, is : could we reasonably expect to get samples like the one we did get if, in fact, $p = 1/2$?

Certainly it is possible to do so. The point is that while every outcome from 0 to 20 is possible, those near 10 are not at all unexpected and the farther we get from 10 the more unexpected (i.e. improbable) they become. This line of argument may be made quantitative by listing the possible outcomes that are more probable (if $p = 1/2$) than that actually obtained, in this instance, 6, 7, 8, 9, 10, 11, 12, 13, 14 and calculating the probability of getting one of these outcomes.

By direct calculation or using suitable tables, this probability turns out to be .958 . It appears, then, that we would seldom get an outcome as far away from 10 as is 15, if $p = \frac{1}{2}$ . The probability of doing so is $1 - .958 = .042$ . Our actual sample, therefore, belongs to a set of samples which is quite improbable if $p = \frac{1}{2}$ and we therefore conclude that to cling to the notion that $p = \frac{1}{2}$ is quite unreasonable. It will be said, in these circumstances, that the outcome, 15, is <u>significantly</u> different from 10, meaning that we cannot reasonably ascribe the discrepancy between these two numbers solely to the errors that arise in sampling.

The probability on which this conclusion is based is the probability of getting a discrepancy as large as or larger than the one observed. It is usually given the symbol P . Smallness of P leads to the pronouncement <u>significant</u>, i.e. there is more here than error. <u>Smallness</u> is, of course, arbitrary.

This procedure could be laid out in another way if we decide in advance what we will consider to be a small probability in this context. Let us say that we decide that any probability P less than .05 will be considered "small". Then, even before the sample

is drawn, we can draw up a list of possible outcomes and attach to each the conclusion that will be reached if this outcome is actually observed. In this example, the list could be drawn up in the following manner.

| Possible outcome | P | Conclusion |
|---|---|---|
| 10 | 1 | NS |
| 9 or 11 | .814 | NS |
| 8 or 12 | .504 | NS |
| 7 or 13 | .264 | NS |
| 6 or 14 | .116 | NS |
| 5 or 15 | .042 | S |
| 4 or 16 | .012 | S |
| 3 or 17 | .002 | S |
| 2 or 18 | .000 | S |
| 1 or 19 | .000 | S |
| 0 or 20 | .000 | S |

The 5% cut-off point is called the level of significance, usually symbolized $\alpha$ .

Approached in this way, the test of significance looks to be the same sort of thing as acceptance sampling. There is, though, a deep and important difference. In acceptance sampling we accept a consignment when our rule fails to reject it. This is not a

conclusion, it is a decision. In the test of significance, failure to attain significance in no way implies the acceptance of anything or a proof of anything. In this example, a sample of 14 would be judged not significant, that is, sampling errors might reasonably have led to this outcome if $p = \frac{1}{2}$, but this does not warrant the conclusion that $p = \frac{1}{2}$.

Another example.

Suppose, in the earlier example, a sample of 100 were taken, yielding a count of 60 voters who will vote for A and 40 who will vote against A. Is this inconsistent with the (hypothetical) proportion $p = \frac{1}{2}$?

This example differs from the earlier one only in the arithmetic involved in the calculation of P. Invoking the normal approximation to the binomial distribution, $\frac{y - 50}{5}$ is N(0,1) and P is 2 Prob$(y \geq 60)$, i.e. 2 Prob$\left(\frac{y - 50}{5} \geq \frac{60 - 50}{5} = 2\right) = .046$ from the normal table, just below the 5% level of significance.

The correction for continuity would lead to $\frac{59.5 - 50}{5} = 1.9$, which leads to $P = .057$.

Certain formalities have grown up around the simple notion of the test of significance, which is, after all, only a sensible

precaution to help us avoid explanations for a difference that error alone may well have produced. It is sometimes said that the test of significance tests the hypothesis that errors alone produced the difference in question. Sometimes this hypothesis is called the null hypothesis and it may even be given a symbol, $H_o$. In the examples, it might be said that we are testing the hypothesis $H_o : p = \frac{1}{2}$, by investigating the difference between $\frac{1}{2}$ and the sample proportion.

Tests of significance based on the normal distribution; confidence intervals; estimation of population parameters.

One of the centrally important problems of statistics may be set up as follows. We have a population (here regarded as infinite), the description of which involves certain unknown constants (mean, variance, perhaps others) which will be called parameters of the population. A sample is to be drawn from the population with a view to estimating the values of these parameters. For obvious reasons, we must insist that the sample shall be drawn randomly. This granted, we face the question : what combinations of the observed values will furnish the required estimates? There are many answers, some better than others. Population parameters will be symbolized by Greek letters and sample estimators of them by the corresponding Latin letters. (For historical reasons, there will be a few exceptions.)

Estimation of the mean of the population.

A population (infinite), described by a statistical variable $x$, has mean $\mu$ and variance $\sigma^2$. A sample of $n$ observations is to be drawn randomly. Designate the values which will be obtained as $x_1, x_2, \ldots x_n$. Then, we may regard the $x_i$ as independent random variables, each distributed according to the distribution of the population. Therefore, each of the $x_i$ has mean $Ex = \mu$ and variance $E(x-\mu)^2 = \sigma^2$.

The average of the sample, $\bar{x} = \frac{1}{n} \sum x_i$, is then a random variable with mean

$$E\bar{x} = \frac{1}{n} \sum E\, x_i = \frac{1}{n} \cdot n\, \mu = \mu \quad \text{and variance}$$

$$\text{Var}\ \bar{x} = \frac{1}{n^2} \sum \text{Var}\ x_i = \frac{1}{n^2} \cdot n\, \sigma^2 = \frac{\sigma^2}{n}.$$

Thus $\bar{x}$ can be treated as a single observation drawn randomly from a distribution with mean $\mu$ and variance $\sigma^2/n$.

Clearly, if the sample size $n$ is large, the variance $\sigma^2/n$ is small and the distribution of $\bar{x}$ is closely clustered about $\mu$. The probability of getting a value of $\bar{x}$ differing appreciably from $\mu$ is therefore small. In this sense (which is called consistency), $\bar{x}$ is a usable estimator of $\mu$. Because $E\bar{x} = \mu$, the estimator $\bar{x}$

of $\mu$ is said to be unbiased. The sample average has other desirable and important properties, as an estimator of $\mu$, which will not be discussed here.

These facts, together with the central limit theorem, provide an important, powerful, statistical instrument, based on the fact that $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is approximately $N(0,1)$. If the population itself is normal, it follows that the distribution of $\bar{x}$ is normal and the approximation becomes an exact statement.

An example. A population has mean $\mu$ (whose value is not known) and variance $\sigma^2$ (known). On the basis of a sample, $x_1, x_2, \ldots x_n$ we wish to check whether the mean could reasonably be taken to be $\mu_o$ (given).

We know that $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is $N(0,1)$. Our question can therefore be framed : if we put $\mu = \mu_o$, is the resulting number $\frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$ the sort of number that could reasonably come randomly from $N(0,1)$ ? If it is not, that is, if $\frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$ differs from zero by so much as to be highly improbable, we conclude that $\mu \neq \mu_o$. We therefore ask the question : is $\frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$ significantly different from zero? Reference to the normal table yields the value of $P$ on which to base our answer.

We can also approach the question, as before, by choosing the level of significance, that is, that value of $P$ such that any smaller value will lead to the pronouncement significant. If, for example, we choose $\alpha = .05$, the normal table yields the statement

$P\left(\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} > 1.96\right) = .05$, or, equivalently, the probability is .95 that $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ lies between $-1.96$ and $1.96$. Hence, when $\mu$ is replaced by $\mu_o$, if this inequality is violated, $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ would be judged significantly different from zero.

This inequality may be arranged in another form.

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \, \sigma/\sqrt{n} \quad .$$

So arranged, we get an interval within which any value $\mu_o$ would be judged not significantly different from $\bar{x}$. Hence, in this sense, it is a not unreasonable possible value of $\mu$. This interval is called a 95% confidence interval for $\mu$. It could be called an interval estimate of $\mu$, but it need not be and will not be here.

The stipulation of known variance, which is crucial to the arguments used in this example, would usually not be satisfied in practice. We must therefore ask how to proceed in this example if the population variance is not known. One possible approach seems obvious : use the sample to estimate the population variance and investigate how to use it in place of the actual population

variance. If we call the estimator $s^2$ (however it may be calculated), we might think of replacing $\sigma$ by $s$ in $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$. This raises what appears to be a question of extraordinary difficulty. The ratio $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ will no longer behave according to $N(0,1)$. How, then, does it behave?

It is a remarkable stroke of luck that this question can be answered simply (in a mathematical sense) and that the answer is simple.

Estimation of the variance of the population.

It will be supposed here that the population is $N(\mu, \sigma^2)$, $\mu$ and $\sigma$ not known. Randomly chosen samples of size $n$ will be designated $x_1, x_2, \ldots x_n$.

Two simple remarks.

1. Any computation we may make with the values of a sample may be spoken of as a transformation.
2. Each observation on $N(\mu, \sigma^2)$, $x_i$ say, may be given a structure $x_i = \mu + e_i$, where the $e_i$ (which may reasonably be called errors) come randomly from $N(0, \sigma^2)$. $\mu$ and the e's cannot be known, but the x's, after the sample is taken, are known numbers.

Think first of samples of 2 observations, $x_1, x_2$ . Let us transform them by writing

$$y_1 = x_1 + x_2$$

$$y_2 = x_1 - x_2$$

Then, writing $x_i = \mu + e_i$ ,

$$y_1 = 2\mu + e_1 + e_2$$

$$y_2 = \quad e_1 - e_2$$

The effect of the transformation is to put everything in the observations concerning $\mu$ into $y_1$ , leaving $y_2$ to deal with error only. We can see that $y_1$ and $y_2$ are both normal, both with variance $2\sigma^2$ and $Ey_1 = 2\mu$ , $Ey_2 = 0$ .

It may appear simpler to change the transformation slightly, so that $y_1 = \frac{x_1 + x_2}{2}$ (i.e. $\bar{x}$) = $\mu$ + error, but in fact a change of a different kind makes for greater simplicity. This change is one that gives $y_1$ and $y_2$ the same <u>variances</u> as the original observations. Thus, we will have

$$y_1 = \frac{1}{\sqrt{2}} x_1 + \frac{1}{\sqrt{2}} x_2$$

$$y_2 = \frac{1}{\sqrt{2}} x_1 - \frac{1}{\sqrt{2}} x_2$$

There will be many occasions for writing such transformations and

it is expedient to adopt a stripped-down way of writing them. For example,

| | $x_1$ | $x_2$ |
|---|---|---|
| $y_1$ | $\frac{1}{\sqrt{2}}$ | $\frac{1}{\sqrt{2}}$ |
| $y_2$ | $\frac{1}{\sqrt{2}}$ | $-\frac{1}{\sqrt{2}}$ |

or, better still,

| | $x_1$ | $x_2$ | divisor |
|---|---|---|---|
| $y_1$ | 1 | 1 | $\sqrt{2}$ |
| $y_2$ | 1 | -1 | $\sqrt{2}$ |

Degrees of freedom.

It may be said that a sample of 2 observations has 2 degrees of freedom, meaning by this that the sample values, $x_1$, $x_2$ say, could be plotted as a point with respect to a pair of axes and that this point may lie anywhere in some portion of the space (i.e. it is not constrained to lie on any curve). In the same sense, the transformed sample has 2 degrees of freedom. One of these degrees of freedom has been assigned to the task of gathering up everything the sample has to say about the value of the population mean $\mu$ ; the other degree of freedom then displays the contribution of error only.

The orthogonal linear transformation.

The reason why the transformation that was used succeeds in dividing the degrees of freedom into two sets with quite different properties rests on a special feature of the transformation

called orthogonality. If any linear transformation of the x's into the y's is written:

$$y_1 = px_1 + qx_2 \qquad p, q, r, s \text{ numbers},$$

$$y_2 = rx_1 + sx_2 \quad ,$$

then, if $pr + qs = 0$ , the transformation is said to be orthogonal. We also imposed another condition, to arrange that $\text{Var } y_1 = \text{Var } y_2 = \sigma^2$ that $p^2 + q^2 = 1$ , $r^2 + s^2 = 1$ . Henceforth, when we speak of an orthogonal transformation, we shall mean a linear transformation satisfying both kinds of conditions, i.e. sum of products = 0 , sum of squares = 1 .

In putting the orthogonal transformation to use, $y_1$ was especially designated to exhibit the contribution of $\mu$ , which amounts to arranging that $y_1$ , is a multiple of $\bar{x}$ . This amounts to choosing $p = q$ . It follows then from orthogonality that $r + s = 0$ and that, in consequence, $y_2$ can reflect no contribution from $\mu$ .

It can be verified that when $p, q, r, s$ are chosen so that $pr + qs = 0$ , $p^2 + q^2 = 1$ , $r^2 + s^2 = 1$ , it follows that $pq + rs = 0$ , $p^2 + r^2 = 1$ , $q^2 + s^2 = 1$ , and that $y_1^2 + y_2^2 = x_1^2 + x_2^2$

The extension to cope with samples larger than 2 proceeds along

the same lines; the sum of products between each pair of rows = 0 and the sum of squares in each row = 1 .

The reason why orthogonal transformations play an important role in the discussion of samples from normal populations is found in a theorem which states that an orthogonal transformation changes a set of independent normal variables, each with variance $\sigma^2$ (the x's) into a set of independent normal variables, each with variance $\sigma^2$ (the y's). Only the means are changed.

Returning now to the question of estimating a population variance : let $x_1, x_2, \ldots x_n$ represent a sample to be chosen randomly from $N(\mu, \sigma^2)$. Then, $x_i = \mu + e_i$ , with $e_i$ : $N(0, \sigma^2)$ . The following orthogonal transformation will separate the error contributions from that of $\mu$ .

| | $x_1$ | $x_2$ . . . | $x_n$ | divisor |
|---|---|---|---|---|
| $y_1$ | 1 | 1 | 1 | $\frac{}{\sqrt{n}}$ |
| $y_2$ | | | | |
| . | | orthogonal | | |
| . | | | | |
| $y_n$ | | | | |

Then, $y_2, y_3 \ldots y_n$ reflect error only. Thus, only (n-1) of

the $n$ degrees of freedom in the sample can be brought to bear on the estimation of error. The way in which these y's may be used to estimate the error variance is probably not obvious, but at least the following facts are easily perceived. For each $y_i$, $i = 2...n$, $Ey_i = 0$, $\text{Var } y_i = \sigma^2$. Hence each $y_i^2$, $i = 2...n$, is a random variable with $Ey_i^2 = \text{Var } y_i = \sigma^2$. Therefore, $y_2^2 + y_3^2 + \cdots y_n^2$ is a random variable with mean $(n-1)\sigma^2$ and $\dfrac{y_2^2 + y_3^2 + \cdots y_n^2}{n-1}$ is a random variable with mean $\sigma^2$. If this quantity is to be used to estimate $\sigma^2$, it will be unbiased. Other important and desirable features of this estimator will not be discussed here.

Lastly, $y_2^2 + y_3^2 + \ldots + y_n^2 = \sum_{i=1}^{n} x_i^2 - y_1^2$

$$= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= \sum (x_i - \bar{x})^2 .$$

The symbol $s^2$ will be given to this estimator of $\sigma^2$.

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \right]$$

$$= \frac{1}{n-1} \left[ \Sigma x_i^2 - n \bar{x}^2 \right] .$$

The divisor, $n - 1$, used to make $s^2$ an unbiased estimator of $\sigma^2$, is seen to be the number of degrees of freedom in the sample which can be devoted exclusively to displaying errors.

The $\chi^2$ distributions

Let $z_1, z_2, \ldots z_p$ be independent standardized normal variables (i.e. each $z_i$ is $N(0,1)$). The sum of their squares, $z_1^2 + z_2^2 + \ldots z_p^2$, is a random variable denoted $\chi^2_{(p)}$, Chi-square with $p$ degrees of freedom. Its probability distribution is known and tabulated.

Refer back to the transformation. $y_2, y_3, \ldots y_n$ are independent normal variables with mean zero and variance $\sigma^2$. Therefore $\frac{y_2}{\sigma}, \frac{y_3}{\sigma}, \ldots \frac{y_n}{\sigma}$ are independent standardized normal variables. It follows that $\frac{1}{\sigma^2}(y_2^2 + y_3^2 + \ldots + y_n^2) = \chi^2_{(n-1)}$ and $\frac{(n-1)s^2}{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{\sigma^2} = \chi^2_{(n-1)}$. The probability distribution of the estimator $s^2$ is therefore known in terms of $\sigma^2$, the actual population variance.

Return now to the question raised earlier : what happens to the standardized normal variable $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ when $\sigma$ is replaced by a sample estimate $\sqrt{s^2} = s$ ? For convenience let us square this ratio and replace $\sigma^2$ by $s^2$, getting

$$\frac{n(\bar{x} - \mu)^2}{s^2} \quad .$$

From the transformation, $y_1$ is a normal variable, mean $\sqrt{n}\,\mu$ and variance $\sigma^2$. Hence $\frac{y_1 - \sqrt{n}\,\mu}{\sigma}$ is $N(0,1)$ and $\left(\frac{y_1 - \sqrt{n}\,\mu}{\sigma}\right)^2$ is $\chi^2_{(1)}$.

Also, $y_1 = \sqrt{n}\,\bar{x}$, so we have

$$\frac{n(\bar{x} - \mu)^2}{\sigma^2} \text{ is } \chi^2_{(1)}$$

Therefore, $\frac{n(\bar{x} - \mu)^2}{s^2} = \frac{n(\bar{x} - \mu)^2/\sigma^2}{s^2/\sigma^2} = \frac{\chi^2_{(1)}}{\chi^2_{(n-1)}/(n-1)}$.

Furthermore, the transformation tells us that these two $\chi^2$'s are independent. The ratio $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ is thus seen to be the ratio of two independent random variables, the numerator $N(0,1)$, the denominator $\sqrt{\chi^2_{(n-1)}/(n-1)}$

The t distributions

The ratio of two independent random variables $\frac{N(0,1)}{\sqrt{\chi^2_{(p)}/p}}$ is given the symbol $t$ or $t_{(p)}$. Its distribution is known and tabulated.

It follows that, when we replace $\sigma$ by $s$ in the ratio $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ the only change required in our procedure is to refer to the t - table instead of the normal table, in performing tests of significance or calculating confidence limits.

The t - distribution is used to check a hypothetical mean $\mu_o$, by substituting $\mu_o$ for $\mu$ in the expression for t and asking if the resulting number could reasonably have come randomly from a t - distribution. Because the t - distribution is symmetrical, this leads to a symmetrical interval about zero, within which we would pronounce the t - values "reasonable" or, if we prefer to deal with the complementary "unreasonable" set of t - values, to two equal tails of the t - distribution.

A $\chi^2$ test of significance.

The $\chi^2$ - distribution provides an instrument for checking a hypothetical variance $\sigma_o^2$. If we substitute $\sigma_o^2$ for $\sigma^2$ in the expression for $\chi^2$, we get $\frac{\sum(x_i - \bar{x})^2}{\sigma_o^2}$ and ask if this number could reasonably come randomly from a $\chi^2$ distribution.

The situation here differs in one respect from that encountered in testing a hypothetical mean. The $\chi^2$ distribution is not symmetrical and neither are the circumstances in which it is used. In virtually all actual circumstances in which we ask : is $\sigma^2 = \sigma_o^2$, the only other possibility that exists is that $\sigma^2 > \sigma_o^2$ and $\frac{\sum(x_i - \bar{x})^2}{\sigma_o^2}$ would differ from $\chi^2$ through being too big. The probability P is therefore calculated from the upper tail of the

$\chi^2$ - distribution. If this probability turns out to be unreasonably small, we conclude that $\sigma^2 > \sigma_o^2$.

Instances arise in which $\frac{\sum(x_i - \bar{x})^2}{\sigma_o^2}$ turns out to be significantly small. It would seldom happen that this finding would lead to the conclusion $\sigma^2 < \sigma_o^2$. Usually it would be suspected that some essential condition had been neglected, for example, a failure to employ randomness where it was needed. Refer to exercise 15.

The $\chi^2$ - and t - distributions both depend on the number of degrees of freedom used in calculating an estimate of variance. Each number of degrees of freedom requires its own set of tabulated values. For this reason, these distributions are tabulated more coarsely than is the normal distribution.

For large numbers of degrees of freedom (more than 30 or 40), the t - distribution approaches the normal so closely that it is sufficient to use the normal table instead of the t - table.

The F distributions

Let $\chi^2_{(p)}$ and $\chi^2_{(q)}$ be two independent random variables, each with a chi-square distribution. The ratio $\frac{\chi^2_{(p)}/p}{\chi^2_{(q)}/q}$, always given the symbol $F_{(p,q)}$, has a known and tabulated distribution.

It is called the analysis of variance distribution, for reasons that will appear subsequently. Observe that $F_{(1,q)} = t^2_{(q)}$ .

This distribution could be put to use in a situation where we have samples from two populations, from which we calculate estimates of the two population variances. We calculate $s_1^2$ from a sample of $n_1$ observations, estimating $\sigma_1^2$ and $s_2^2$ from a sample of $n_2$ observations, estimating $\sigma_2^2$ .

Then, $\dfrac{(n_1 - 1)s_1^2}{\sigma_1^2}$ is $\chi^2_{(n_1 - 1)}$

$\dfrac{(n_2 - 1)s_2^2}{\sigma_2^2}$ is $\chi^2_{(n_2 - 1)}$

Therefore, $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is $F_{(n_1-1,\ n_2-1)}$ .

This fact would be useful if we wished to ask (as we occasionally do) : is $\sigma_1^2 = \sigma_2^2$ ? If so, $\dfrac{s_1^2}{s_2^2}$ is $F_{(n_1-1,\ n_2-1)}$ and we could ask, then, could this number, $\dfrac{s_1^2}{s_2^2}$ , reasonably come randomly from the F distribution? The details of this test will be discussed later, when it will be contrasted with another quite different use of the F - distribution.

Further on two-sample problems.

Usually, samples are drawn from two populations to study the difference between the mean of the populations. Indeed, in most situations of this sort, we can be assured on prior grounds that the variances are equal and the question of equality of variances does not arise.

Suppose, then, that one sample, $x_{11}, x_{12}, \ldots x_{1n}$ comes randomly from population 1, $N(\mu_1, \sigma^2)$ and another $x_{21}, x_{22}, \ldots x_{2n}$ comes from population 2, $N(\mu_2, \sigma^2)$.

We can write $x_{1\alpha} = \mu_1 + e_{1\alpha}$

$$x_{2\alpha} = \mu_2 + e_{2\alpha}$$

where the e's may be thought of as coming randomly from $N(0, \sigma^2)$.

Set up the following orthogonal transformation.

| | $x_{11}$ | $x_{12}$ | ... | $x_{1n}$ | $x_{21}$ | $x_{22}$ | ... | $x_{2n}$ | divisor |
|---|---|---|---|---|---|---|---|---|---|
| $y_1$ | 1 | 1 | ... | 1 | 1 | 1 | ... | 1 | $\sqrt{2n}$ |
| $y_2$ | 1 | 1 | ... | 1 | -1 | -1 | ... | -1 | $\sqrt{2n}$ |
| $y_3$ | | | | | | | | | |
| . | | | | | | | | | |
| . | orthogonal | | | | | | | | |
| . | | | | | | | | | |
| $y_{2n}$ | | | | | | | | | |

Then, $y_1 = \sqrt{\frac{n}{2}}\ (\mu_1 + \mu_2) + \text{errors}$

$y_2 = \sqrt{\frac{n}{2}}\ (\mu_1 - \mu_2) + \text{errors}$

$y_3, y_4, \ldots y_{2n}$ contain errors only.

Therefore, $y_3^2 + y_4^2 + \ldots + y_{2n}^2 = \sigma^2\ \chi^2_{(2n-2)}$

and $s^2 = \frac{1}{2n-2}\ (y_3^2 + y_4^2 \ldots + y_{2n}^2)$ is an unbiased estimator of $\sigma^2$.

If $\mu_1 = \mu_2$, $y_2^2 = \sigma^2\ \chi^2_{(1)}$ and $\frac{y_2^2}{s^2} = F_{(1,\ 2n-2)}$

and $\frac{y_2}{s} = t_{(2n-2)}$. Either of these ratios may be used to test whether $y_2$ contains more than error, that is, if $y_2$ contains a non-zero contribution from $\mu_1 - \mu_2$.

Computing rules may be devised in the same way as before. For convenience, write $T_i$ for $\sum_{\alpha=1}^{n} x_{i\alpha}$ and G (grand total) for $T_1 + T_2$. Then,

$y_1 = \frac{G}{\sqrt{2n}}$, $y_2 = \frac{T_1 - T_2}{\sqrt{2n}}$, from which $y_1^2 + y_2^2$ can be calculated directly from the observations.

Then, $y_3^2 + y_4^2 + \ldots + y_{2n}^2 = \sum_{i=1}^{2} \sum_{\alpha=1}^{n} x_{i\alpha}^2 - y_1^2 - y_2^2$

$$= \sum \sum x_{i\alpha}^2 - \frac{(T_1 + T_2)^2}{2n} - \frac{(T_1 - T_2)^2}{2n}$$

$$= \sum x_{1\alpha}^2 - \frac{T_1^2}{n} + \sum x_{2\alpha}^2 - \frac{T_2^2}{n}$$

$$= \sum (x_{1\alpha} - \bar{x}_1)^2 + \sum (x_{2\alpha} - \bar{x}_2)^2 \quad .$$

Another way of calculating $y_2^2$, and the way in which it is reached, are useful and instructive. As far as $y_1$ and $y_2$ are concerned, the transformation can be written

| | $T_1$ | $T_2$ | divisor |
|---|---|---|---|
| $\sqrt{n}y_1$ | 1 | 1 | $\sqrt{2}$ |
| $\sqrt{n}y_2$ | 1 | -1 | $\sqrt{2}$ |

Therefore $n(y_1^2 + y_2^2) = T_1^2 + T_2^2$ and

$$y_2^2 = \frac{1}{n}(T_1^2 + T_2^2) - \frac{G^2}{2n} \quad .$$

It is convenient to carry out these calculations and record the outcome in an analysis of variance table.

| source of variation | degrees of freedom | sums of squares |
|---|---|---|
| between samples | 1 | $y_2^2 = \frac{1}{n}(T_1^2 + T_2^2) - \frac{G^2}{2n}$ |
| within samples | 2n - 2 | by subtraction |
| "total" | 2n - 1 | $y_2^2 + y_3^2 \ldots + y_{2n}^2 = \sum\sum x_{i\alpha}^2 - \frac{G^2}{2n}$ |

One or two other columns may be added if desired.

| | mean squares | F |
|---|---|---|
| between | $y_2^2/1$ | $y_2^2/s^2$ |
| within | $(y_3^2 + \ldots + y_{2n}^2)/(2n - 2)$ which is $s^2$ | |

The component $y_1^2$ has not been entered in the table, although its value has been used in the calculations as $G^2/_{2n}$ (sometimes called the correction factor, C.F.). It is not entered because $y_1$, which displays the contribution of $\mu_1 + \mu_2$, is never the object of inquiry when several samples are taken. The object of such sampling is always to estimate differences or contrasts. Among any number $m$ of objects, there are $m - 1$ algebraically independent differences and $m - 1$ degrees of freedom are required to display these differences. Applying this way of looking at sets of observations to our two-sample problem : we have $2n$ observations, hence $2n - 1$ d.f. are required to display the differences among them. This is the "total" d.f. listed in the table. These differences are of two kinds, between samples, which can reflect the value of $\mu_1 - \mu_2$ and within the individual samples, which cannot reflect the value of $\mu_1 - \mu_2$. There are two samples, hence 1 d.f. for the difference between them and within each sample of $n$ observations there are $n - 1$ d.f. for differences. There are altogether, then, $2(n - 1)$ d.f. for within samples differences.

If the two samples are not equal in size, the same approach

(left as an exercise) yields the analysis of variance table:

| source | d.f. | s.s. |
|---|---|---|
| between samples | 1 | $y_2^2 = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} - \frac{G^2}{n_1 + n_2}$ |
| within samples | $n_1 + n_2 - 2$ | by subtraction |
| total | $n_1 + n_2 - 1$ | $y_2^2 + y_3^2 \ldots + y_{n_1+n_2}^2 = \sum \sum x_{i\alpha}^2 - \frac{G^2}{n_1+n_2}$ |

Cause and Effect

Problems involving two samples can arise in all sorts of ways. The most important of these comes from experimental situations, in which we introduce changes in a "causal" system, hoping to detect resulting changes in an "effect" system. An experiment, then, is carried out to reach conclusions about cause and effect.

If a change from $C_1$ to $C_2$ in a causal system is accompanied by a change from $E_1$ to $E_2$ in an effect system, it seems reasonable to conclude that the change $C_1 \rightarrow C_2$ causes the change $E_1 \rightarrow E_2$ provided we know that

(1) it happens every time it is tried;

(2) nothing else is responsible for the observed change $E_1 \rightarrow E_2$.

These conditions are seen to be impossible to meet. The second one implies that we must know all the agents which could bring about the change $E_1 \rightarrow E_2$, (an unlikely situation) and that we have prevented them from doing so (experimental control).

Further, the presence of error is sure to interfere in both conditions (1) and (2). Statistics and in particular that portion of it called Design of Experiments, is concerned with finding some accommodation between what we would like to do and what we can do.

These notions may be crystallized somewhat by thinking about an actual experimental situation. Let us say that we wish to compare two diets, $D_0$, a "basic" diet and $D_1$, the basic diet plus some additive, by feeding them to rats and recording the gains in weight.

I. Several rats will be assigned to each diet, not just one to each. This may be thought of as an attempt to meet stipulation (1) or, to say the same thing in another way, by having several rats on the same diet we are able to perceive, through the differences among them, the contributions of error.

II. Say we have decided to put $n$ rats on diet $D_0$ and $n$ rats on diet $D_1$. We must therefore assemble $2n$ experimental animals. These animals are sure to differ from one another with respect to the responses they will furnish when they are given the diets. Usually, some of the reasons for these differences will be known or suspected beforehand, (the most important ones, we hope); others, we may

suppose, we cannot even guess.

The planning of an experiment, generally speaking, is made up of two parts; first, we make use of our knowledge of the reasons for (possibly large) differences to arrange that these differences shall not be allowed to enter into the comparisons we wish to study or into the definition of error; second, to deal with sources of differences which we cannot control, usually because we are not aware of them, we insist that they go randomly into the experiment, by selecting randomly the animals to go into each category of the experiment. This random allocation is an attempt to avoid coming in conflict with stipulation (2), inasmuch as it avoids bias and has the further consequence that differences of the sort we are randomizing out are made to behave as if they come randomly from a distribution of errors and hence can be dealt with by probability theory.

III. Usually "external" controls are needed. In the case of our rats, we would insist that they all be housed in the same way, with uniform light and heat, with feeding conditions suitably controlled and so on. Controls of this sort depend entirely on the knowledge of the experimenter. They must be taken for

granted here. In any event, they are dictated by stipulation (2).

Returning to the example, let us pursue the simplest case by supposing that there are no discernible reasons why one rat should give a different response from another. We then allocate the $2n$ rats randomly into two groups of $n$, feed $D_0$ to one group, $D_1$ to the other. We then have two samples arising out of the experiment.

The objective in performing this experiment is, of course, to study $\bar{x}_1 - \bar{x}_2$, the estimator of $\mu_1 - \mu_2$, to interpret it in terms of cause and effect and, perhaps, to construct a confidence interval for this difference, As a rather important detail in the course of the study, we would seek assurance, through a test of significance, that the difference $\bar{x}_1 - \bar{x}_2$ could not reasonably be accounted for on the basis of error only. We do this simply to avoid devising explanations or reaching conclusions about a difference $\mu_1 - \mu_2$ which may well be non-existent.

The standard formulae which come out of this discussion are:

(a) The test of significance;

$$F_{(1,\ 2n-2)} = \frac{\frac{n}{2}(\bar{x}_1-\bar{x}_2)^2}{s^2} \quad , \text{ where } \quad s^2 = \frac{\sum(x_{1\alpha}-\bar{x}_1)^2 + \sum(x_{2\alpha}-\bar{x}_2)^2}{2n - 2}$$

or an equivalent formula

$$t_{(2n-2)} = \frac{\sqrt{\frac{n}{2}}(\bar{x}_1-\bar{x}_2)}{s} \quad ;$$

(b) confidence limits for $\mu_1 - \mu_2$ at level $1 - \alpha$ are given by

$$\bar{x}_1 - \bar{x}_2 \pm t_{(\alpha/2)} s\sqrt{\frac{2}{n}}$$

Extension to three or more samples.

If we think of adding a third diet $D_2$, made up, let us say, by starting with the basic diet $D_0$ and adding twice as much of the additive as was used to form $D_1$, the experiment would yield three samples. The analysis of variance dictated by the structure of the experiment must be, if $n$ rats are assigned to each diet:

| source | d.f. |
|---|---|
| among samples | 2 |
| within samples (error) | 3n - 3 |
| total | 3n - 1 |

The only new question to arise here concerns the 2 d.f. among samples. Evidently two components in the transformation must be assigned to these contrasts. How should they be specified?

There is no single correct answer to this question. It depends on the nature of the differences introduced into the causal system and the reasons why they were selected. A particular instance will be pursued here and discussed later.

Call the observations $x_{i\alpha}$ , $i = 1,2,3$, $\alpha = 1,2,\ldots n$ .

Then $x_{i\alpha} = \mu_i + e_{i\alpha}$ .

| | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1n}$ | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2n}$ | $x_{31}$ | $x_{32}$ | $\cdots$ | $x_{3n}$ | div. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_1$ | 1 | 1 | ... | 1 | 1 | 1 | ... | 1 | 1 | 1 | ... | 1 | $\sqrt{3n}$ |
| $y_2$ | -1 | -1 | ... | -1 | 0 | 0 | ... | 0 | 1 | 1 | ... | 1 | $\sqrt{2n}$ |
| $y_3$ | -1 | -1 | ... | -1 | 2 | 2 | ... | 2 | -1 | -1 | ... | -1 | $\sqrt{6n}$ |
| $y_4$ | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | |
| . | orthogonal | | | | | | | | | | | | |
| . | | | | | | | | | | | | | |
| $y_{3n}$ | | | | | | | | | | | | | |

Then, $y_2 = \sqrt{\frac{n}{2}}(\mu_3 - \mu_1) + \text{error}$

$y_3 = \sqrt{\frac{n}{6}}(-\mu_1 + 2\mu_2 - \mu_3) + \text{error}$

$y_4, y_5, \ldots y_{3n} = \text{error only.}$

It should be clear that $y_4^2 + y_5^2 + \ldots + y_{3n}^2 = \sum_{i=1}^{3} \sum_{\alpha=1}^{n} (x_{i\alpha} - \bar{x}_i)^2$ and that $s^2$ , the estimator of $\sigma^2$ , is this s.s. divided by $3(n-1)$ . We can therefore calculate $y_2$ and $y_3$ and test them against error to find out if either of them contains more than error. Unless

neither of them does so, we would have evidence that the three diets yield results that are not all the same. The question still to be settled concerns the specific nature of the differences produced by the three diets.

If the three means, $\mu_1$, $\mu_2$, $\mu_3$ were known, we could in this instance plot them against the amounts of additive used to make up the diets, i.e. 0, 1, 2 . Any question about differences among the diets is a question about the nature of the curve joining the three points of the graph.

The following facts come from simple geometry. If $-\mu_1 + 2\mu_2 - \mu_3 = 0$ , the points $(0,\mu_1)$ , $(1,\mu_2)$ , $(2,\mu_3)$ lie on a straight line and the slope of this line is proportional to $\mu_3 - \mu_1$ . The interpretation of $y_2$ and $y_3$ should now be clear. If $y_3$ appears to contain nothing but error, gain in weight depends linearly on the amount of additive. If this proves to be the case, it becomes necessary to test $y_2$ to find out if this line has a zero slope (the diets yield the same gains in weight) or a slope different from zero (the diets yield genuinely different gains in weight).

If $y_3$ is significantly different from zero, there is no occasion to test $y_2$ . (interpretation?)

Of course, the multipliers in $y_2$ and $y_3$ were chosen to facilitate this study of the curve relating gain in weight with amount of additive. This particular choice of multipliers depends on the averages $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ being based on the same numbers of observations and abscissae 0, 1, 2 being equally spaced. When the averages are based on the same numbers of observations, it will at least be possible to use suitable and meaningful multipliers. In any truly experimental situation, the object of the exercise will be to perceive certain well defined contrasts which will, in turn dictate the choice of the multipliers. Often the set of contrasts form a hierarchy in which the order in which the contrasts are scrutinized is important.

If several more diets are included in the experiment, more components are needed to describe the curve and the choice of the proper multipliers to use would be dependent on the structure of the diets.

Orthogonal Experiments

An experiment which can be analysed by means of an orthogonal transformation is said to be an orthogonal experiment. "Analysis" in this context, means the extraction from the observations of those contrasts which the experiment was designed to study.

As an example of non-orthogonality, if an experiment of the sort we have been discussing were carried out using different numbers of rats on

the various diets, it would in general prove to be impossible to study meaningful contrasts among diets by means of an orthogonal transformation. Orthogonality in the design, then, requires equal samples. There are a few special instances in which this condition is not essential to the use of an orthogonal transformation, but even so, equal sample sizes are important and should always be planned for.

To sum up the computations for several samples, for the case of 3 samples:

| source | d.f. | s.s. |
|---|---|---|
| among samples | 2 | $y_2^2 + y_3^2 = \frac{1}{n} \sum_{i=1}^{3} T_i^2 - \frac{G^2}{3n}$ |
| within samples | 3(n - 1) | $y_4^2 + \ldots + y_{3n}^2$ (computed by subtraction) |
| total | 3n - 1 | $\sum_i \sum_\alpha x_{i\alpha}^2 - \frac{G^2}{3n}$ |

This preliminary table should always be calculated, even though it does not exhibit everything we want. It does not require that choices of multipliers have been made, because the among samples sum of squares is the same for all choices. We would, however, proceed to separate the among samples s.s. into $y_2^2$ and $y_3^2$ using the transformation.

If the samples are of different sizes, $n_1$, $n_2$, $n_3$, the among samples s.s. is given by

$$\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - \frac{G^2}{n_1 + n_2 + n_3}$$

and the total s.s. by

$$\sum \sum x_{i\alpha}^2 - \frac{G^2}{n_1 + n_2 + n_3}$$

Observe that when the calculations in an analysis of variance table are carried out in the patterns shown, only one kind of computation is used; certain numbers are squared and divided by something, the divisor is always the number of observations added to produce the number that is squared.

The experimental arrangements that have been discussed so far are called completely randomized designs.

Note. The use of gain in weight, i.e. the difference between the final and the initial weight of the animal, may not be the best combination of these numbers to use; this question will be raised later.

The details of F - tests.

Looking back at one of the examples, we find

$$y_2 = \sqrt{\frac{n}{2}}\,(\mu_3 - \mu_1) + \text{error}$$

$y_4$, $y_5$, ... $y_{3n}$ display error only.

That is, $y_4, y_5 \ldots y_{3n}$ are independent normal variables, each with mean 0 and variance $\sigma^2$. Hence, $y_4^2 + y_5^4 \ldots + y_{3n} = \sigma^2 \chi^2_{(3n-3)}$

Now, <u>if</u> $\mu_3 - \mu_1 = 0$, $y_2$ is $N(0,\sigma^2)$ and, since it is independent of all the other y's,

$$\frac{y_2^2}{(y_4^2 + y_5^2 + \ldots + y_{3n}^2)/(3n-3)} = F_{(1,\ 3n-3)}$$

Now there is only one way in which this ratio can fail to be an F. If $\mu_3 - \mu_1 \neq 0$, $y_2^2$ is to be expected to be too big and the ratio will be too big to be an F. If the ratio, when it is calculated, turns out to be less than 1, there is nothing to test; but if it proves to be greater than 1, we must ask : is the ratio too large to come randomly from an F - distribution? The test is one - tailed and the F - tables are made up for this test.

When the F - test is used to decide $\overset{\text{whether}}{\wedge}$ two population variances are equal, using sample estimates $s_1^2$ and $s_2^2$, we use the fact that $\frac{s_1^2}{s_2^2} = F$ when the variances are equal. In this case, there are no prior grounds for asserting that the mean of $s_1^2$ is either greater or less than the mean of $s_2^2$ and the question we must ask is : is our calculated ratio too different from 1 to be an F - value? The test is two - tailed.

To use the F - table, which was made up for the one - tailed test, put the larger of $s_1^2$ and $s_2^2$ in the numerator, the other in the denominator. Then look up this number in the F - table and get the probability P . Lastly, double the value of P .

Further on the structure of experiments.

Return to the experiment in which diets $D_0$ and $D_1$ are to be tested on rats and add to what was assumed earlier the knowledge or belief that males and females may respond rather differently to the diets. If this is so, to ignore the sexes of the animals in allocating them to the diets amounts to allowing a source of (possibly large) systematic variation to run through the comparisons the experiment is intended to display, inflating the error sum of squares and distorting the comparison between the diets. A better plan would take account of the possibility of a systematic difference between males and females and arrange to display the difference between the diets within each of the sexes in an orthogonal manner. This can be accomplished by assigning randomly the same number n of males and females to each of the diets.

In one respect, we have nothing new here. We have a completely randomized experiment with 4 samples, each of n observations. If

we label the samples $D_0S_1$, $D_1S_1$, $D_0S_2$, $D_1S_2$ and $T_{D\ S}$ etc. represent the totals of the samples, we have an analysis of variance table

| source | d.f. | s.s. |
|---|---|---|
| among samples | 3 | $\frac{1}{n}\sum T_{DS}^2 - \frac{G^2}{4n}$ |
| within samples | 4(n-1) | by subtraction |
| total | 4 n-1 | $\sum x^2 - \frac{G^2}{4n}$ |

The aspect of the experiment that <u>is</u> new is the way in which we must study the 3 d.f. among the samples. This is dictated by the manner in which the samples were caused to be different. We envisaged two sources of variation, which will be called <u>factors</u>, each of which was introduced at two levels, each level of one factor tested in combination with each level of the other factor. Arrangements of this sort are called <u>factorial</u> arrangements and the experiment may be called a factorial experiment.

Analysis of the 3 d.f. among samples amounts to sorting out the variation caused by changing each of the factors and, first of all, checking whether it is possible to do so.

Consider the portion of an orthogonal transformation that has to do with the 3 d.f. among samples. It can be laid out as follows.

| | $T_{D_0S_1}$ | $T_{D_1S_1}$ | $T_{D_0S_2}$ | $T_{D_1S_2}$ | div. |
|---|---|---|---|---|---|
| $y_2$ | 1 | 1 | -1 | -1 | $\sqrt{4n}$ |
| $y_3$ | 1 | -1 | 1 | -1 | $\sqrt{4n}$ |
| $y_4$ | 1 | -1 | -1 | 1 | $\sqrt{4n}$ |

$y_2$ displays the difference between $S_1$ and $S_2$ averaged over the diets; $y_3$ displays the difference between the diets, averaged over the sexes. This averaging is entirely proper, provided the differences being averaged are estimates of the same population-difference. We need assurance, for example, that the difference caused by changing the diet is the same for males as for females before we combine them in an average.

The component $y_4$ is seen to exhibit the difference between these differences we wish to average. If $y_4$ is not significantly different from zero, we can proceed with the averaging and the components $y_2$ and $y_3$ are proper and meaningful and may be tested if we wish. On the other hand, if $y_4$ is significantly different from zero, $y_2$ and $y_3$ represent averages that should not be calculated and certainly should not be tested. If $y_4$ is significantly

different from zero, it will be said that sexes and diets interact and, in any event, $y_4$ is called an interaction component. Note that the coefficients of $y_4$ in the transformation can be formed by multiplying the corresponding coefficients of $y_2$ and $y_3$. For this reason, the interaction of sexes and diets is symbolized sex x diet.

Our procedure, then, is to calculate $y_2^2$, $y_3^2$, $y_4^2$ and record their values in the analysis of variance table. They may be calculated from the transformation, but it is simpler to calculate them according to the following scheme.

| | $S_1$ | $S_2$ | |
|---|---|---|---|
| $D_0$ | $T_{D_0S_1}$ | $T_{D_0S_2}$ | $T_{D_0}$ |
| $D_1$ | $T_{D_1S_1}$ | $T_{D_1S_2}$ | $T_{D_1}$ |
| | $T_{S_1}$ | $T_{S_2}$ | $G$ |

| source | d.f. | s.s. |
|---|---|---|
| sexes | 1 | $\frac{1}{2n}\left(T_{S_1}^2 + T_{S_2}^2\right) - \frac{G^2}{4n}$ |
| diets | 1 | $\frac{1}{2n}\left(T_{D_0}^2 + T_{D_1}^2\right) - \frac{G^2}{4n}$ |
| sexes x diets | 1 | by subtraction |
| "Total" | 3 | $\frac{1}{n}\left(T_{D_0S_1}^2 + T_{D_0S_2}^2 + T_{D_1S_1}^2 + T_{D_1S_2}^2\right) - \frac{G^2}{4n}$ |

First, test the sex x diet interaction for significance; if not significant, proceed to study sex and diet differences; if significant, there is no occasion to test sex and diet differences (which are often called main effects). Note that, when we have demonstrated that a real sex x diet exists (i.e. $y_4$ is significantly different from zero), we have at the same time confirmed that the sexes do respond differently and the diets do yield different gains in weight.

The notion of interaction has been encountered for the first time in this example. It is a centrally important concept in virtually all experimentation.

The definition and estimation of error

It may appear, from the examples discussed so far, that there is only one way of displaying error, by arranging that several individuals are assigned to each sample (i.e. treated alike in all respects) so that differences among them must reflect error only. However, this is not the only way of providing for the display of error; indeed it is not usually the most effective way.

To perceive the point of view at issue here, think again of an experiment involving experimental animals, in which two or more treatments of a sort that can be applied to the skin of the animal are to be compared.

We could, of course, divide the animals into groups, each

group to receive one of the treatments. This would be a completely randomized experiment and differences between animals, in their responses to the treatments, would contribute to error.

Another possibility may exist here. If the several treatments can be applied to different parts of the same animal, the differences we want to study are perceived within animals and differences between animals do not affect them. Neither will they affect the error, properly defined.

Suppose only two treatments, $t_1$ and $t_2$, are to be compared by selecting 2 sites on each of $n$ animals, and on each assign randomly $t_1$ to one of the sites and $t_2$ to the other. There will be $2n$ observations, with a total of $2n - 1$ d.f. to study. We may, to start with, think first of separating these d.f. as follows:

| | d.f. |
|---|---|
| among animals | $n - 1$ |
| within animals | $n$ |
| total | $2n - 1$ |

Now within animals, we have 1 d.f. to display the difference between

the two treatments. The remaining $n - 1$ d.f. can then presumably reflect error only.

The final analysis of variance table will then be :

| | d.f. |
|---|---|
| among animals | $n - 1$ |
| between treatments | 1 |
| error | $n - 1$ |
| total | $2n - 1$ |

The computation of these sums of squares should be obvious.

It may be useful to write an orthogonal transformation for this arrangement. Use $n = 4$ and call the animals $A_1$ , $A_2$ , $A_3$ , $A_4$ .

| | $A_1$ | | $A_2$ | | $A_3$ | | $A_4$ | | div. |
|---|---|---|---|---|---|---|---|---|---|
| | $t_1$ | $t_2$ | $t_1$ | $t_2$ | $t_1$ | $t_2$ | $t_1$ | $t_2$ | |
| $y_1$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | $\sqrt{8}$ |
| $y_2$ | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | $\sqrt{8}$ |
| $y_3$ | 1 | 1 | -1 | -1 | 1 | 1 | -1 | -1 | $\sqrt{8}$ |
| $y_4$ | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | $\sqrt{8}$ |
| $y_5$ | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | $\sqrt{8}$ |
| $y_6$ | 1 | -1 | 1 | -1 | -1 | 1 | -1 | 1 | $\sqrt{8}$ |
| $y_7$ | 1 | -1 | -1 | 1 | 1 | -1 | -1 | 1 | $\sqrt{8}$ |
| $y_8$ | 1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | $\sqrt{8}$ |

Then, $y_2$ , $y_3$ , $y_4$ are the 3 d.f. for animals, $y_5$ is the 1 d.f. for treatment and $y_6$ , $y_7$ , $y_8$ display error only. These facts may be checked by letting $x_{ij}$ stand for the observation on animal $A_i$ and treatment $t_j$ and noting that

$x_{ij} = \alpha_i + \mu_j + e_{ij}$ .

Even though there has been no repetition of observations to furnish an estimate of the error variance, something has been repeated, namely, the difference we are investigating. It has been exhibited within each animal; the extent to which these differences differ among themselves is a measure of the error with which the treatment difference is estimated. Inspection of the error components verifies that they are made up of differences among the treatment differences.

Since each animal furnishes a complete display of the difference under inquiry (or, more generally, of all the contrasts under inquiry), it is customary to say that each animal furnishes a replication. It is also said that each animal constitutes a block (the word comes from agriculture). Blocking or stratification implies an attempt to group experimental material into sets that are more uniform than is the material as a whole, within which contrasts may be studied, exposed to smaller errors than would otherwise be possible. Replication

implies blocking. (The term replication is best used only in this sense. It is best not to use the word to describe the repetitions in a completely randomized experiment).

Reference to the transformation shows that the error components have the structure of an interaction between replications and treatments. There may be some temptation here to speak of replications as a factor and to regard this experimental arrangement as an instance of a factorial experiment. It is best not to do so. The "factor" replications is, by definition, one which cannot interact with treatments. It is the responsibility of the experimenter to see to it that this condition is satisfied.

Experimental arrangements of this kind are called randomized block arrangements. It has been implied that each block shall be "big" enough to contain all the contrasts under study and the "treatments" are allocated randomly to the experimental units of each block. If there are $t$ treatments and $b$ blocks, the analysis of variance table is:

| | d.f. |
|---|---|
| blocks (or replications) | $b - 1$ |
| treatments | $t - 1$ |
| blocks x treatments (error) | $(b - 1)(t - 1)$ |
| total | $bt - 1$ |

We would, of course, inquire in more detail into the $t - 1$ d.f. for treatments.

Paired comparisons.

The particular instance of a randomized block experiment in which there are only two treatments can be approached in what seems to be a different manner. From each block we can calculate the difference $d_i = x_{1i} - x_{2i}$ and treat these numbers as a sample from a single population. $\bar{d} = \bar{x}_1 - \bar{x}_2$ evidently estimates $\mu_1 - \mu_2$. If we wish to test whether $\mu_1 - \mu_2 = 0$, or to calculate confidence limits for $\mu_1 - \mu_2$, we would calculate $s^2 = \frac{1}{n-1} \sum (d_i - \bar{d})^2$ and $t_{(n-1)} = \frac{\bar{d} - (\mu_1 - \mu_2)}{s/\sqrt{n}}$.

It is important to perceive the distinction between the paired comparison experiment and the completely randomized two sample experiment.

More on factorial experiments.

The term factorial arrangement will be construed as applying only to the structure of the treatment differences put into the causal system by introducing one factor at $p$ levels, another at $q$ levels, another at $r$ levels, and so on. The set of different treatment combinations formed by combining each level of each factor with each level of every other factor will be called a $p \times q \times r \ldots$ factorial

arrangement. The number of different treatments generated in this way is $p \times q \times r \ldots$ .

Usually, the levels of some of the factors can be specified numerically and the results of the experiment can be plotted as a set of graphs. Analysis of the results then becomes a comparison of a number of curves to find out in what ways they are alike and in what ways they differ, with a view to deciding how much reduction or simplification is warranted.

As an example, think of a $4 \times 2$ factorial experiment, carried out in a randomized block pattern with $r$ replications. A preliminary analysis of variance table reads:

| | d.f. |
|---|---|
| replications | $r - 1$ |
| treatments | 7 |
| error | $7(r - 1)$ |

Our concern is with the interpretation of the 7 d.f. for treatments.

Let A be one of the factors, with levels $a_0$ , $a_1$ , $a_2$ , $a_3$ , which can properly be plotted as 0 , 1 , 2 , 3 . Let B be the other factor, with levels $b_0$ , $b_1$ . We might, for example, be making up 8 different fertilizers by starting with some basic fertilizer, adding nitrogen to it in amounts 0 , 1 , 2 , 3 units and adding phosphorous in amounts 0 , 1 units. There will be 3 d.f. for

comparing the 4 levels of A , 1 d.f. for comparing the 2 levels of B and $3 \times 1 = 3$ d.f. for A × B interactions. The computations corresponding to this separation of d.f. may be carried out from the following table.

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ | |
|---|---|---|---|---|---|
| $b_0$ | $T_{a_0b_0}$ | $T_{a_1b_0}$ | $T_{a_2b_0}$ | $T_{a_3b_0}$ | $T_{b_0}$ |
| $b_1$ | $T_{a_0b_1}$ | $T_{a_1b_1}$ | $T_{a_2b_1}$ | $T_{a_3b_1}$ | $T_{b_1}$ |
| | $T_{a_0}$ | $T_{a_1}$ | $T_{a_2}$ | $T_{a_3}$ | G |

$T_{a_ib_j}$ is the sum of the r observations on treatment $a_ib_j$ .

| | d.f. | s.s. |
|---|---|---|
| A | 3 | $\frac{1}{2r}\sum T^2_{a_i} - \frac{G^2}{8r}$ |
| B | 1 | $\frac{1}{4r}\sum T^2_{b_j} - \frac{G^2}{8r}$ |
| A × B | 3 | by subtraction |
| total treatments | 7 | $\frac{1}{r}\sum T^2_{a_ib_j} - \frac{G^2}{8r}$ |

At this point, we may still not be in a position to interpret the results of the experiment, but if it should happen that the A × B

interaction s.s. is significantly large, we could reasonably be assured of the existence of a genuine interaction between the A and B factors. In this event, any attempt to interpret the A and B s.s. is unwarranted.

If the A × B interaction s.s., with its 3 d.f. is not significantly large, this fact by itself does not give adequate assurance that there is no genuine A × B interaction and we are obliged to look more closely into the interaction s.s. This is best carried out by means of a suitable orthogonal transformation.

| | $T_{a_0b_0}$ | $T_{a_1b_0}$ | $T_{a_2b_0}$ | $T_{a_3b_0}$ | $T_{a_0b_1}$ | $T_{a_1b_1}$ | $T_{a_2b_1}$ | $T_{a_3b_1}$ | div. |
|---|---|---|---|---|---|---|---|---|---|
| A lin | -3 | -1 | 1 | 3 | -3 | -1 | 1 | 3 | $\sqrt{20r}$ |
| A quad | -1 | 1 | 1 | -1 | -1 | 1 | 1 | -1 | $\sqrt{8r}$ |
| A cub | -1 | 3 | -3 | 1 | -1 | 3 | -3 | 1 | $\sqrt{20r}$ |
| B | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | $\sqrt{8r}$ |
| B × $A_\ell$ | 3 | 1 | -1 | -3 | -3 | -1 | 1 | 3 | $\sqrt{20r}$ |
| B × $A_q$ | 1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | $\sqrt{8r}$ |
| B × $A_c$ | 1 | -3 | 3 | -1 | -1 | 3 | -3 | 1 | $\sqrt{20r}$ |

The multipliers in $A_\ell$, $A_q$, $A_c$ come from a table of orthogonal polynomial values.

If one or more of the components $B \times A_\ell$ , $B \times A_q$ , $B \times A_c$ , are significantly different from zero, we would be obliged to conclude that a genuine $A \times B$ interaction exists and the first four components in the transformation would not be useful. On the other hand, if each of the interaction components is not significant, the average A and B curves become meaningful. In particular, the components $A_\ell$ , $A_q$ , $A_c$ are useful in deciding what kind of curve is adequate.

Further on interactions.

In the factorial examples we have been discussing, absence of interaction is reflected in parallelism of a number of curves. This is the usual case, but there is one situation in which a rather different meaning should be attached to the notion of interaction. To make an example, suppose we wish to compare two different supplements to a basic diet, both intended to produce the same response and in the same way. To be specific, let us say that they are to be compared by feeding diets made up by adding to the basic diet one unit and two units of each of the additives.

If the additives are in fact equivalent, they should produce the same response curves. If one of them is more concentrated than the

other, they should produce the same response at the zero level (whether this is actually tested or not), a difference should show up at level 1 and twice this difference should be found at level 2.

Clearly the response curves here cannot be expected to be parallel; they are lines radiating from a point. If two different lines are produced, we would conclude that the additives are not equally concentrated (or potent). If we find that (difference at level 2) - 2 (difference at level 1) reflects nothing but error, we would conclude that the additives differ only in potency, but if this difference of differences looms large, we could not account for the difference between additives solely on the grounds of potency and would be obliged to conclude that there is a qualitative difference as well.

If we call the additives $A_1$ and $A_2$ and the levels $\ell_1 = 1$ and $\ell_2 = 2$, an orthogonal transformation which displays a suitable interaction for this situation is

| | $A_1$ | | $A_2$ | |
|---|---|---|---|---|
| | $\ell_1$ | $\ell_2$ | $\ell_1$ | $\ell_2$ |
| levels | -1 | 1 | -1 | 1 |
| additives | -1 | -2 | 1 | 2 |
| $\ell \times a$ | 2 | -1 | -2 | 1 |

A thorough discussion of experiments of this kind, in which we encounter the interaction of quantity and quality, requires rather elaborate methods. Mostly they are encountered in biological assay.

In some instances, particularly those in which we are not working too close to the zero level of application, an excellent way of circumventing the difficulty we have been discussing is to plot response against log (level). This transformation has the effect of changing a set of curves ∧ radiating from the same point at zero level into a set of parallel curves, when there are no interactions. The notion of interaction then reverts to its usual meaning, lack of parallelism.

If we intend to employ this approach, we should think of the possibility of choosing levels in geometric progression (e.g. 1, 2, 4... units) in order that their logarithms are in arithmetic progression. Analysis of the results is thereby simplified because orthogonal polynomials may be used to study the shape of the response curves.

Confounding - incomplete blocks.

The randomized block experimental arrangement is a sort of reference plan against which all others may be assessed. Probably other arrangements would be used only rarely, except for the fact that

physical conditions frequently impose limitations which stand in the way of making full use of this plan. One limitation often encountered arises because useful and sensible blocks are not big enough to contain a complete replication. This situation may be encountered for all sorts of reasons. The factorial experiment is often the culprit, because it becomes very large very fast, as we add factors and levels.

We are sometimes obliged, then, to use several blocks for each replication. The blocks are then said to be <u>incomplete</u>. All that will be attempted here is to pursue the consequence of the incomplete block in a special diminutive example.

Think of a $2 \times 2 \times 2$ (or $2^3$) factorial arrangement which is to be tested in $r$ replications.

Call the factors A , at levels $a_1$ and $a_2$ ,
B , at levels $b_1$ and $b_2$ ,
C , at levels $c_1$ and $c_2$ .

There are then 8 treatment combinations, which may be symbolized $a_i b_j c_k$ , $i$ , $j$ , $k = 1$ , $2$ .

The preliminary analysis of variance table would read

| | <u>d.f.</u> |
|---|---|
| reps | $r - 1$ |
| treatments | 7 |
| error | $7(r - 1)$ |

The 7 treatments d.f. would be separated into components corresponding to the main effects and interactions as displayed in the following transformation. Let the symbol $(a_i b_j c_k)$ stand for the total of the observations on the treatment $a_i b_j c_k$ .

| | $(a_1b_1c_1)$ | $(a_1b_1c_2)$ | $(a_1b_2c_1)$ | $(a_1b_2c_2)$ | $(a_2b_1c_1)$ | $(a_2b_1c_2)$ | $(a_2b_2c_1)$ | $(a_2b_2c_2)$ |
|---|---|---|---|---|---|---|---|---|
| A | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 |
| B | -1 | -1 | 1 | 1 | -1 | -1 | 1 | 1 |
| C | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 |
| AB | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 |
| BC | 1 | -1 | -1 | 1 | 1 | -1 | -1 | 1 |
| CA | 1 | -1 | 1 | -1 | -1 | 1 | -1 | 1 |
| ABC | -1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |

Suppose now that suitable blocks are available of a size that will accommodate only 4 treatments. It therefore becomes necessary to use 2 blocks in each replication and we have to decide which treatments to put in each block.

The consequences of any particular decision are easily perceived by referring to the transformation. Suppose, for example, that the first four treatments listed in the transformation are put into one

block and the others into the other block and that the same allocation to blocks is used in each replication. It is clear, then, that the main effect A gathers up also, in each replication, the difference between the two blocks in that replication and that this component exhibits, in addition to whatever main effect A there may be, the difference between one set of r blocks and the other set of r blocks. In this circumstance, it is said that the main effect A is confounded with blocks (or block differences).

Inspection of the other components in the transformation shows that they are not confounded with blocks, inasmuch as their coefficients add to zero within each block.

Evidently we can choose an allocation into blocks that will confound any one of the components with blocks and leave the others free from block differences. It seems obvious also that an injudicious allocation may confound several components with blocks.

Turning now to the analysis and interpretation of a confounded experiment, let us pursue the instance in which the component A is confounded.

There are two possibilities to be considered. If block differences are thought to be not too large and may be thought of as simply con-

tributing another source of error in the A comparison, we may think of part of the analysis in the form

| | d.f. |
|---|---|
| reps | r - 1 |
| A | 1 |
| reps × A | r - 1 (error term for the A comparison) |

The rest of the analysis would take the form

| | d.f. |
|---|---|
| treatments | 6 |
| B | 1 |
| C | 1 |
| A B | 1 |
| B C | 1 |
| C A | 1 |
| A B C | 1 |
| reps × treatments | 6(r - 1) (error term for this portion of the analysis). |

If, as is often the case, block differences are quite large and we are content to sacrifice a component which is confounded with them, we need only calculate

| | d.f. |
|---|---|
| blocks | 2r - 1 |
| treatments | 6 |
| error | 6(r - 1) |

The arithmetic is best carried out in a succession of steps which correspond to the following analysis of variance tables.

| | d.f. | s.s. |
|---|---|---|
| among blocks | $2r - 1$ | $\frac{1}{4} \sum (\text{block totals})^2 - \frac{G^2}{8r}$ |
| within blocks | $6r$ | by subtraction |
| total | $8r - 1$ | $\sum (\text{observations})^2 - \frac{G^2}{8r}$ |

The sums of squares corresponding to the factorial components may be calculated from the transformation, but it may be more convenient to obtain them from a set of two - way tables like the following.

| | | |
|---|---|---|
| $T_{a_1 b_1}$ | $T_{a_1 b_2}$ | $T_{a_1}$ |
| $T_{a_2 b_1}$ | $T_{a_2 b_2}$ | $T_{a_2}$ |
| $T_{b_1}$ | $T_{b_2}$ | $G$ |

Analysis of variance calculations on this table yield

| | d.f. | s.s. |
|---|---|---|
| A | 1 | $\frac{1}{4r}(T_{a_1}^2 + T_{a_2}^2) - \frac{G^2}{8r}$ |
| B | 1 | $\frac{1}{4r}(T_{b_1}^2 + T_{b_2}^2) - \frac{G^2}{8r}$ |
| A × B | 1 | by subtraction |
| "total" | 3 | $\frac{1}{2r} \sum T_{a_i b_j}^2 - \frac{G^2}{8r}$ |

The two other similar tables yield the s.s. for A (which need not be calculated again), C and A × C and B , C and B × C .

The s.s. for A , B , C , A × B , B × C , C × A , together with the s.s. for A × B × C , make up the "treatment" s.s. with 7 d.f., which is calculated from

$$\frac{1}{r} \sum T^2_{a_i b_j c_k} - \frac{G^2}{8r} \quad .$$

We can get the A B C s.s. by calculating this s.s. and subtracting all the other s.s. from it.

The computations may now be organized and finished off according to the following pattern.

| | d.f. | | d.f. | s.s. |
|---|---|---|---|---|
| among blocks | 2r - 1 | reps | r - 1 | $\frac{1}{8} \sum (\text{rep. totals})^2 - \frac{G^2}{8r}$ |
| | | A | 1 | already calculated |
| | | error | r - 1 | by subtraction |
| within blocks | 6r | B | 1 | |
| | | C | 1 | |
| | | A B | 1 | already calculated |
| | | B C | 1 | |
| | | C A | 1 | |
| | | A B C | 1 | |
| | | error | 6(r - 1) | by subtraction |

The split plot arrangement.

It may happen, in a factorial experiment, that one factor is necessarily confounded with a certain source of error, but the other factor need not be. For example, a factor A might be several levels of a drug administered to rats. Comparisons among levels of the drug inevitably involve differences among rats. An experiment to study these comparisons might be arranged in a completely randomized, randomized block or other pattern. Let us say, now, that the response to be studied is the threshold at which the animal reacts to an electrical stimulus and that several types of stimulus are to be used. Types of electrical stimulus, then, constitute several levels of a factor B , each of which may be applied to the same rat. Comparisons among the levels of B , then, need not be exposed to differences among rats.

Let us say, to have something definite to discuss, that factor A has 3 levels of drug, each of which is administered to n rats in a completely randomized design and that factor B consists of 4 types of electrical stimulus, each of which is applied to each rat. The experiment will then furnish 12 n observations.

The natural separation of degrees of freedom, to start with, is among and within rats.

| | d.f. |
|---|---|
| among rats | 3n - 1 |
| within rats | 9n |
| total | 12n - 1 |

Indeed, the sums of squares corresponding to this separation ought to be computed.

The s.s. among rats is then separated into two parts : among levels of $A$ , with 2 d.f. and within levels of $A$ , with 3n - 3 d.f., which is the error s.s. for testing $A$ .

The s.s. within rats separates into three parts : among levels of $B$ , with 3 d.f.; $A \times B$ with 6 d.f.; error with $9(n - 1)$ d.f., which is the error s.s. for testing $A \times B$ and $B$ .

No doubt, in an actual experiment, we would want to inquire further into the $A \times B$ , $B$ and $A$ sums of squares.

Experiments with this kind of structure are called split plot, after an agricultural prototype. The structure of the split plot experiment is seen to be identical with that of the incomplete block experiment discussed earlier. In the split-plot situation, though, it is inevitable that a main effect shall be confounded with the additional source of error.

Regression Analysis

Regression theory is concerned with relationships among variables, some or all of which are statistical variables in the sense that they have frequency distributions. This is, of course, the kind of question we have been discussing in connection with experiments. Indeed, the most important uses of regression theory are in the analysis of experiments in which orthogonality is lacking. It is used in other contexts as well, with appropriate reservations and precautions.

Presumably it would be generally agreed that weights of men are, in some sense, related to their heights, that tall men tend to be heavier than short men, even though we may know a short man, $S$, who is heavier than a tall man, $T$, whom we know. It is not at all clear, though, how such a qualitative statement can be made quantitative because some or even all the members of the population will not satisfy any relation we might specify. We are here in a different position from the physicist who says that the distance a spring stretches depends on the force applied to it and writes $D = kF$. He means by this that $D$ and $F$ are strictly related by this equation; that every time a force $F_0$ is applied,

the stretch $D_0$ and no other, always results. He may grant, though, that this relation may appear not to hold exactly, because of the intervention of errors of measurement.

In our situation, we know that several men, all the same height, are likely to show a considerable variety of weights. Let us restate the question : given a group of men of given height, what can we say about their weights? We can then envisage a population of men, all the same height, of which our group of weights constitutes a sample, presumably randomly chosen. We can then think of such a population of weights corresponding to every height. If these populations change from one height to another, this reflects a dependence of some sort of weight on height.

It is to be emphasized that the question, as posed here, is a conditional one : given height, what can we say about weight? Height, here, is not a statistical variable. It need not come randomly from a frequency distribution, that is, there is no error connected with it. If the sampling is such that heights do come randomly from a population of heights, this fact is irrelevant to the question being asked. Weight, on the other hand, is a statistical variable.

Questions of this sort are usually put in a much more definite and restricted form. If we use $x$ to denote height and $y$ weight, we might ask : how does the mean of $y$ , <u>for given</u> $x$ (i.e. $E(y/x)$) vary with $x$ ? In particular, does it vary linearly with $x$ and, if so, how do we estimate this linear relation from a set of observations? We will start with the second question, <u>assuming</u> an affirmative answer to the first and later take up the first question.

We assume, then, that $E(y/x) = \beta_o + \beta_1 x$ , $\beta_o$ , $\beta_1$ unknown. It follows that, for any recorded height $x_\alpha$ and the corresponding observed weight $y_\alpha$ ,

$y_\alpha = \beta_o + \beta_1 x_\alpha + e_\alpha$ , $e_\alpha$ an error.

We will suppose that the errors we encounter in a sample come randomly from a single error distribution with variance $\sigma^2$ (unknown). This is a strong assumption. It says that the error variance is the same for all values of $x$ . It is the same assumption as was made in the discussion of experiments, where it can usually be supported, but in other circumstances it can easily be violated.

<u>The sample</u>.

From what has been said, we could in principle settle on several heights (x-values) and then sample one or more weights

(y-values) corresponding to these heights. This might be troublesome to carry out in this instance, but in some others it would be the simplest and most natural way of getting the sample. An alternative plan is to choose randomly a sample of men and record the height and weight of each. Under either scheme, we come up with a number (say $n$) of pairs of numbers $x_\alpha$, $y_\alpha$, $\alpha = 1, 2 \ldots n$.

It is customary to speak of $x$ as the independent variable, the fixed (i.e. non-statistical) variable or the selector variable and of $y$ as the dependent or statistical variable.

Estimation

Having a suitably chosen sample and assuming that $E(y/x) = \beta_o + \beta_1 x$, we proceed to estimate the values of $\beta_o$, $\beta_1$ and $\sigma^2$.

Let $b_o$ and $b_1$ represent numbers to be calculated from the sample to estimate $\beta_o$ and $\beta_1$ and write

$$Y = b_o + b_1 x .$$

Then evidently $Y$ is an estimator of $E(y/x)$ for any given $x$.

Whatever values may be given to $b_o$ and $b_1$, a value $Y_\alpha$ may be calculated corresponding to each $x_\alpha$ in the sample : $Y_\alpha = b_o + b_1 x_\alpha$. Obviously we want the $Y_\alpha$ values to be as

close to the $y_\alpha$ values as possible. One way of accomplishing this is to choose the b's , i.e. the Y's , so that $\sum_{\alpha=1}^{n} (y_\alpha - Y_\alpha)^2$ is made as small as possible. This is the "principle of least squares". It is demonstrable that this principle, in terms of the results it yields, is "best" in terms of more basic principles of estimation.

The choice of $b_o$ and $b_1$ to minimize $\sum(y_\alpha - Y_\alpha)^2$ leads to two conditions to be satisfied by them.

$$\sum(y_\alpha - Y_\alpha) = 0$$

$$\sum(y_\alpha - Y_\alpha)x_\alpha = 0 \ .$$

Substituting for the $Y_\alpha$ in these equations,

$$n\, b_o + b_1 \sum x_\alpha = \sum y_\alpha$$

$$b_o \sum x_\alpha + b_1 \sum x_\alpha^2 = \sum x_\alpha y_\alpha \ .$$

These equations are called <u>normal</u> equations (no connection with the normal distribution). The first of these equations may be written

$$b_o + b_1 \bar{x} = \bar{y}$$

which asserts that the regression line must pass through the average point $(\bar{x}, \bar{y})$ and that its equation may be written

$$Y = \bar{y} + b_1(x - \bar{x}) \ .$$

Indeed, with this knowledge, if we had set out to fit the line in

the form $Y = b_o' + b_1(x - \bar{x})$ , the normal equations would read

$$n\, b_o' + b_1 \sum (x_\alpha - \bar{x}) = \sum y_\alpha$$

$$b_o' \sum (x_\alpha - \bar{x}) + b_1 \sum (x_\alpha - \bar{x})^2 = \sum y_\alpha (x_\alpha - \bar{x}) \ .$$

Since $\sum (x_\alpha - \bar{x}) = 0$ , the equations virtually solve themselves.

$$n\, b_o' = \sum y_\alpha \ ,$$

$$b_1 \sum (x_\alpha - \bar{x})^2 = \sum y_\alpha (x_\alpha - \bar{x}) \ .$$

$$b_o' = \bar{y} \ , \quad b_1 = \frac{\sum y_\alpha (x_\alpha - \bar{x})}{\sum (x_\alpha - \bar{x})^2} \ .$$

There are alternate formulae that can be useful when calculating $b_1$.

$$\sum y_\alpha (x_\alpha - \bar{x}) = \sum y_\alpha x_\alpha - \frac{(\sum x_\alpha)(\sum y_\alpha)}{n} \ .$$

$$\sum (x_\alpha - \bar{x})^2 = \sum x_\alpha^2 - \frac{(\sum x_\alpha)^2}{n} \ .$$

Note also that

$$\sum y_\alpha (x_\alpha - \bar{x}) = \sum (y_\alpha - \bar{y})(x_\alpha - \bar{x}) \ .$$

The estimation of $\sigma^2$

It seems obvious that if, in fact, $E(y/x) = \beta_o + \beta_1 x$ , $\sum (y_\alpha - Y_\alpha)^2$ reflects error only and therefore can be made the

basis for estimating the error variance. This and other facts come easily out of an appropriate orthogonal transformation.

| | $y_1$ | $y_2 \cdots y_n$ | div. |
|---|---|---|---|
| $z_1$ | 1 | 1 ... 1 | $\sqrt{n}$ |
| $z_2$ | $x_1 - \bar{x}$ | $x_2 - \bar{x} \ldots x_n - \bar{x}$ | $\sqrt{\sum (x_\alpha - \bar{x})^2}$ |
| $z_3$ | | | |
| $z_4$ | | | |
| . | orthogonal | | |
| . | | | |
| $z_n$ | | | |

However the components $z_3, z_4, \ldots z_n$ may be chosen, orthogonality with $z_1$ and $z_2$ ensures that their means must be zero. If, at this point, we introduce the condition that the errors are normally distributed, we see that $z_3^2 + z_4^2 \ldots + z_n^2 = \sigma^2 \chi^2_{(n-2)}$ and that $s^2 = \frac{1}{n-2}(z_3^2 + z_4^2 \ldots + z_n^2)$ estimates $\sigma^2$. Note also that $z_1 = \sqrt{n}\, b_o' = \sqrt{n}\, \bar{y}$ and $z_2 = \sqrt{\sum (x_\alpha - \bar{x})^2}\, b_1$.

Writing $y_\alpha = \beta_o + \beta_1 x_\alpha + e_\alpha$ in the expression for $z_2$,

$$z_2 = \frac{1}{\sqrt{\sum (x_\alpha - \bar{x})^2}} \sum (x_\alpha - \bar{x})(\beta_o + \beta_1 x_\alpha + e_\alpha)$$

$$= \frac{1}{\sqrt{\sum (x_\alpha - \bar{x})^2}} \beta_1 \sum x_\alpha (x_\alpha - \bar{x}) + \text{error}$$

$$= \beta_1 \sqrt{\sum (x_\alpha - \bar{x})^2} + \text{error} \quad .$$

Hence, $E\, z_2 = \sqrt{\sum (x_\alpha - \bar{x})^2}\; E\, b_1 = \sqrt{\sum (x_\alpha - \bar{x})^2}\; \beta_1$ and

$E\, b_1 = \beta_1$ , that is, $b_1$ is an unbiased estimator of $\beta_1$ .

Also, $\text{Var}\, z_2 = \sum (x_\alpha - \bar{x})^2\, \text{Var}\, b_1 = \sigma^2$ and

$$\text{Var}\, b_1 = \frac{\sigma^2}{\sum (x_\alpha - \bar{x})^2}$$

It follows that, if the errors are normally distributed,

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{\sum (x_\alpha - \bar{x})^2}} \quad \text{is} \quad N(0,1) \quad \text{and}$$

$$\frac{b_1 - \beta_1}{s / \sqrt{\sum (x_\alpha - \bar{x})^2}} \quad \text{is} \quad t_{(n-2)} \quad .$$

The obvious computation of the error s.s. is

$$z_3^2 + z_4^2 + \ldots + z_n^2 = \sum y_\alpha^2 - z_1^2 - z_2^2$$

$$= \sum y_\alpha^2 - n\, b_o'^2 - b_1^2 \sum (x_\alpha - \bar{x})^2$$

$$= \sum y_\alpha^2 - \frac{(\sum y_\alpha)^2}{n} - \frac{\left[\sum y_\alpha (x_\alpha - \bar{x})\right]^2}{\sum (x_\alpha - \bar{x})^2}$$

This s.s. can be identified with $\sum(y_\alpha - Y_\alpha)^2$, as is to be expected. This will be done shortly. First, however, there is a fundamental identity of regression analysis, which asserts that

$$\sum(y_\alpha - \bar{y})^2 = \sum(y_\alpha - Y_\alpha)^2 + \sum(Y_\alpha - \bar{y})^2 .$$

This can be checked by writing

$$\sum(y_\alpha - \bar{y})^2 = \sum(y_\alpha - Y_\alpha + Y_\alpha - \bar{y})^2$$

$$= \sum(y_\alpha - Y_\alpha)^2 + \sum(Y_\alpha - \bar{y})^2 + 2\sum(y_\alpha - Y_\alpha)(Y_\alpha - \bar{y})$$

Now, $\sum(y_\alpha - Y_\alpha)(Y_\alpha - \bar{y}) = \sum(y_\alpha - Y_\alpha)Y_\alpha - \bar{y}\sum(y_\alpha - Y_\alpha)$

$$= \sum(y_\alpha - Y_\alpha)(b_o + b_1 x_\alpha) - \bar{y}\sum(y_\alpha - Y_\alpha)$$

$$= b_o\sum(y_\alpha - Y_\alpha) + b_1\sum(y_\alpha - Y_\alpha)x_\alpha - \bar{y}\sum(y_\alpha - Y_\alpha) .$$

Each of these sums has value zero, in virtue of the normal equations.

Now the error s.s., $z_3^2 + z_4^2 + \ldots z_n^2 = \sum(y_\alpha - \bar{y})^2 - b_1^2\sum(x_\alpha - \bar{y})^2$ and, since $Y_\alpha - \bar{y} = b_1(x_\alpha - \bar{x})$ , $\sum(Y_\alpha - \bar{y})^2 = b_1^2\sum(x_\alpha - \bar{x})^2$ , the error s.s. is $\sum(y_\alpha - \bar{y})^2 - \sum(Y_\alpha - \bar{y})^2 = \sum(y_\alpha - Y_\alpha)^2$

It is useful to record the results of the computations in an

analysis of variance table.

| | d.f. | s.s. |
|---|---|---|
| attributable to regression | 1 | $z_2^2 = b_1^2 \sum(x_\alpha - \bar{x})^2 = \left[\sum y_\alpha (x_\alpha - \bar{x})\right]^2 \Big/ \sum(x_\alpha - \bar{x})^2$ |
| deviations from regression | n-2 | $z_3^2 + z_4^2 + \ldots + z_n^2$ , by subtraction |
| total | n-1 | $z_2^2 + z_3^2 + \ldots + z_n^2 = \sum y_\alpha^2 - \frac{(\sum y_\alpha)^2}{n}$ |

We can, if we wish, test $z_2^2$ to see if it contains more than error (i.e. $\beta_1 \neq 0$) by asking if $\frac{\text{s.s. regression}/1}{\text{s.s. deviations}/(n-2)}$ is $F_{(1,\ n-2)}$. This ratio is seen to be the square of the $t_{(n-2)}$ reached earlier, with $\beta_1 = 0$ .

Adequacy of the assumption $E(y/x) = \beta_o + \beta_1 x$ .

The estimation of error from deviations from the fitted regression and the test of significance depend heavily on the adequacy of a linear model. If, in fact, some other function ought to be fitted, the s.s. $\sum(y_\alpha - Y_\alpha)^2$ will contain, in addition to error, systematic departures of the assumed line from the correct function and will therefore be too large compared with error. This can be checked if we can get an estimate of error which does not

depend on any fitting. The possibility of accomplishing this rests on the taking of the sample in a special way. If we settle on a number of different values of $x$ and sample each of the populations so selected several times, differences within these samples will reflect error only. Indeed, we are now in the situation described earlier as a completely randomized experiment, with an analysis of variance that reads:

| | d.f. |
|---|---|
| among samples | k-1 |
| within samples (error) | n-k. |

Now, if we fit a regression $Y = b_0 + b_1 x$, one of the $k-1$ d.f. is used up for the slope of the line, with s.s. $b_1^2 \sum (x_\alpha - \bar{x})^2$ and the residual s.s. with $k-2$ d.f. follows by a subtraction. The computation may be summed up as follows.

| | d.f. | s.s. | | |
|---|---|---|---|---|
| among | k-1 | S.S.A | 1 regression | S.S.R |
| | | | k-2 deviations by subtraction | S.S.D |
| within (error) | n-k | S.S.E. by subtraction | | |
| total | n-1 | | | |

Then, $\frac{S.S.D/(k-2)}{S.S.E/(n-k)} = F_{(k-2,\ n-k)}$ if S.S.D contains only error.

The correlation coefficient

When a linear regression $Y = b_0 + b_1 x$ is found to fit the observations adequately, the coefficient $b_1$ displays the dependence, of $y$ on $x$. The numerical value of $b_1$, in itself, carries no conviction of dependence, because this value depends, among other things, on the units in which $x$ and $y$ are measured. A test of significance is required.

The dependence of $b_1$ on the units of measurement is easily removed by multiplying it by $\frac{\sqrt{\sum (x_\alpha - \bar{x})^2}}{\sqrt{\sum (y_\alpha - \bar{y})^2}}$, a quantity that has no bearing on the question of dependence of $y$ on $x$. The resulting coefficient is called the coefficient of correlation of $x$ and $y$, symbolized always by $r$. Thus:

$$r = b_1 \frac{\sqrt{\sum (x_\alpha - \bar{x})^2}}{\sqrt{\sum (y_\alpha - \bar{y})^2}} = \frac{\sum (y_\alpha - \bar{y})(x_\alpha - \bar{x})}{\sqrt{\sum (y_\alpha - \bar{y})^2 \sum (x_\alpha - \bar{x})^2}}$$

Clearly this coefficient is not needed, in this context, at least. It rests on the same assumptions as does the linear regression and can accomplish no more than the coefficient $b_1$ does. Nevertheless it is often calculated and used, indeed, it is often misused through neglecting to check the linearity of the relation and the constancy

of the variance. Used along with a regression analysis, it has a useful property. From the definition of $r$, we have

$$\sum (Y_\alpha - \bar{y})^2 = b_1^2 \sum (x_\alpha - \bar{x})^2 = r^2 \sum (y_\alpha - \bar{y})^2$$

$$\sum (y_\alpha - Y_\alpha)^2 = (1-r^2) \sum (y_\alpha - \bar{y})^2$$

It is seen that $r^2$ is the fraction of the total variation, measured by $\sum (y_\alpha - \bar{y})^2$, that is accounted for by the regression. These relations also display the fact that $r^2 \leq 1$, a property thought by some to be important.

If $r$ is to be used instead of $b_1$ to detect dependence of one variable on another, it is important that a test of significance be used. The test for $b_1$, based on the ratio $\dfrac{\sum (Y_\alpha - \bar{y})^2}{\sum (y_\alpha - Y_\alpha)^2/(n-2)}$, becomes in terms of $r$, $\dfrac{r^2(n-2)}{1-r^2}$, which is $F_{(1,\, n-2)}$ if $\beta_1 = 0$. Alternatively $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ is $t_{(n-2)}$.

Precision of the estimate of $E(y/x)$.

Sometimes the purpose in fitting a regression of $y$ on $x$ is to estimate $E(y/x)$ for various values of $x$.

To judge the precision of such an estimate, we require the variance of $Y$ .

When the regression equation is in the form $Y = \bar{y} + b_1(x - \bar{x})$ , we know that $\bar{y}$ and $b_1$ are independent, with $\text{Var } \bar{y} = \frac{\sigma^2}{n}$ , $\text{Var } b_1 = \frac{\sigma^2}{\sum (x_\alpha - \bar{x})^2}$ . Therefore,

$$\text{Var } Y = \text{Var } \bar{y} + (x - \bar{x})^2 \text{ Var } b_1$$

$$= \sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_\alpha - \bar{x})^2} \right) .$$

$\sigma^2$ is estimated by $s^2 = \frac{1}{n-2} \sum (y_\alpha - Y_\alpha)^2$ .

It follows that

$$\frac{Y - E(y/x)}{s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_\alpha - \bar{x})^2}}} = t_{(n-2)}$$

We can, if we wish, use this to calculate confidence limits for $E(y/x)$ . They are given by

$$Y \pm t_{(\alpha/2)} \, s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_\alpha - \bar{x})^2}} .$$

Regression with two or more independent variables.

The ideas and the approach used in the discussion of linear regression extend without change to the situation in which several independent

variables are required. If these independent variables are denoted $x_1, x_2, \ldots x_p$, the regression equation is

$$E(y/x_1, x_2, \ldots x_p) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p .$$

The estimating relation is

$$Y = b_o + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p ,$$

where the b's are chosen to minimize $\sum (Y_\alpha - Y_\alpha)^2$. The b's are obtained by solving the set of normal equations

$$n\, b_o + b_1 \sum x_{1\alpha} + b_2 \sum x_{2\alpha} + \ldots + b_p \sum x_{p\alpha} = \sum Y_\alpha$$

$$b_o \sum x_{1\alpha} + b_1 \sum x_{1\alpha}^2 + b_2 \sum x_{1\alpha} x_{2\alpha} + \ldots + b_p \sum x_{1\alpha} x_{p\alpha} = \sum x_{1\alpha} Y_\alpha$$

$$\vdots$$

$$b_o \sum x_{p\alpha} + b_1 \sum x_{p\alpha} x_{1\alpha} + b_2 \sum x_{p\alpha} x_{2\alpha} + \ldots + b_p \sum x_{p\alpha}^2 = \sum x_{p\alpha} Y_\alpha$$

Each $b_i$ is an unbiased estimator of $\beta_i$, normally distributed with variance $\sigma^2 c_{ii}$, where the $c_{ii}$ are numbers calculated from the x - observations. The value of $\sigma^2$ is estimated by $s^2 = \dfrac{\sum (Y_\alpha - Y_\alpha)^2}{n - p - 1}$ and the numerator of $s^2$ can be calculated from the formula

$$\sum (Y_\alpha - Y_\alpha)^2 = \sum Y_\alpha^2 - b_o \sum Y_\alpha - b_1 \sum x_{1\alpha} Y_\alpha - \ldots - b_p \sum x_{p\alpha} Y_\alpha .$$

As a special case, we can take $x_2$ to be $x_1^2$, $x_3$ to be $x_1^3$ so on. This theory, therefore, includes the fitting of curved regressions of the sort $E(y/x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots$ .

The analysis of covariance.

In some of the examples used to illustrate some of the ideas in the planning and conduct of experiments, we spoke of comparing diets by feeding them to animals, taking for granted the existence of some sensible measurement that would reflect differences among the diets.

If all the animals entered the experiment at the same weight (initial weight), no doubt their weights at the end of the experiment (final weights) would be a suitable measure. The equality of the initial weights would be next to impossible to arrange, of course, and in any event it is not particularly desirable to do so. We can and should keep the initial weights within some reasonable range and let us say that we record the initial weight of each animal. Call it $x$. Then, at the conclusion of the experiment, we measure its final weight, $y$.

One might be tempted to analyse the differences, $y - x$, in the

belief that by so doing, variation in the final weights attributable to variation in initial weights would be removed; or perhaps, use $y/x$ . The use of either of these devices rests on an assumption which is rather unlikely to be satisfied. In any event, why make assumptions when the observations can supply information about the manner in which $y$ varies with $x$ ? We have only to plot $y$ against $x$ , separately for each diet. Each diet, then, yields its own graph. These graphs ought to be sensibly straight lines, at least if initial weights have been prevented from varying too widely, and these lines ought to be parallel. [Question: What conclusion if the lines turn out to be not parallel] If they are, differences between lines indicate differences between diets and deviations of observations from their individual lines reflect error.

Clearly we have here a special use of regression theory. It is called the analysis of covariance. The arithmetical aspects of this analysis are not entirely obvious, but they are easily mastered when they are needed.

Observations made by counting.

Methods of analysis discussed up to this point (apart from the discussion of samples from binomial populations) require that the observations be measurements. When they arise in the form of

counts, these methods, while not strictly correct, may still be useful in view of the central limit theorem, in the same way that the normal distribution is useful in calculating binomial probabilities. Some special difficulties can arise, though.

As an example, think of an experiment to compare several methods of planting seedlings. In each plot are planted the same number of seedlings, say $n$, and the observation made on each plot will be the number, say $x$, of seedlings that survive and grow or, equivalently, the proportion $p = \frac{x}{n}$ of survivors.

Each plot may be considered to provide a sample of $n$ observations from a binomial population with some unknown proportion $\pi$ of survivors. If the methods of planting do differ, the value of $\pi$ will vary over the experiment. The proportions $p$ are estimates of the corresponding values of $\pi$.

It is entirely reasonable to treat the observed proportions $p$ as if they are continuous variables and to employ analysis of variance procedures, apart from the fact that the condition of uniform error variance is not met. The variance of $p$ is $\frac{\pi(1 - \pi)}{n}$. Even so, if only values of $\pi$ not too far from $\frac{1}{2}$, say between .3

and .7 occur in the experiment, the error variance will not change enough to do serious damage to the analysis.

When values of $\pi$ close to 0 or 1 do occur, it is a fact that analysing $\sin^{-1}\sqrt{p}$ instead of p greatly extends the range of values of $\pi$ over which there is reasonable uniformity in the error variance.

Other transformations are available to deal with some other situations in which the error variance is not constant.

Goodness-of-fit.

Sometimes the observations take the form of counts of individuals within categories. The records of a hospital might show, for example, that of 10,000 infants born there, 5200 were females and 4800 were males. Pursuing this example a little farther, suppose we ask if this sample really indicates that female births are more frequent than males. This calls for a test of significance, really a binomial test but with the normal approximation to the binomial, we would calculate $\frac{5200 - 5000}{\sqrt{10{,}000 \times \frac{1}{2} \times \frac{1}{2}}} = 4$, which is much too large to have come randomly from the standard normal distribution.

We could frame the question in another way. Let us enter our

observations in a table,

observed table

| | |
|---|---|
| M | 4800 |
| F | 5200 |

10,000

and construct another table showing the numbers we would "expect" when a sample of 10,000 is drawn from a binomial population whose proportion is $\frac{1}{2}$ .

expected table

| | |
|---|---|
| M | 5000 |
| F | 5000 |

10,000

We may now seek some measure to display the overall discrepancy between these two tables. The measure used for this purpose, using O to denote an observed count and E the corresponding expected value, is $\sum \frac{(O-E)^2}{E}$ , added over all the cells. In this example, using this measure we would calculate

$\frac{(4800 - 5000)^2}{5000} + \frac{(5200 - 5000)^2}{5000}$ which yields 16, the square of of the number obtained in the earlier test. Recalling that the square of a standard normal variable is $\chi^2_{(1)}$, we see that we could check the 16 against $\chi^2_{(1)}$ as an alternative to checking the 4 against $N(0,1)$. In either case, the test is approximate because the normal distribution is an approximation to the binomial distribution.

The computation used to arrive at the $\chi^2_{(1)}$ can always be carried out to compare a set of observed counts with a corresponding set of expected values, reached on the basis of some prior question or speculation. The resulting number can be checked against a $\chi^2$ - distribution. The one question that needs scrutiny is the number of degrees of freedom of the $\chi^2$.

One might ask, for example, why the $\chi^2$ in the foregoing example does not have 2 d.f., because two pairs of cells are being contrasted. The answer lies in the fact that one of these contrasts adds nothing to the other and is, indeed, predictable from it. This stems from the fact that the expected table was forced to have the same total as the observed table.

The categories or cells in which we record our counts may themselves

be classified or patterned, corresponding to the objectives in making the observations and the questions we ask may force the expected table to agree with the observed table in several respects. Consider an example.

Suppose that some individuals in a population have been inoculated against some infection and some have not. At some suitable point in time, a sample of individuals is taken and each individual is classified as having been inoculated or not (I or $\overline{I}$) and also as having contracted the disease or not (C or $\overline{C}$). There are then four categories. The numbers of individuals in these categories would reasonably be assembled in a $2 \times 2$ table.

| | I | $\overline{I}$ | |
|---|---|---|---|
| C | a | b | a + b. |
| $\overline{C}$ | c | d | c + d |
| | a+c | b+d | N |

Presumably the reason for this exercise would be to find out if the proportions of individuals contracting the disease are appreciably different for the inoculated and not-inoculated groups. We proceed, then, to set up the expected table, under the supposition that the true proportions are the same for the two groups, and test the discrepancy between the observed and the expected tables.

The proportions we are discussing are $\frac{a}{a+c}$ and $\frac{b}{b+d}$. If they

are, in fact, equal, there is nothing to test and the observed table is the same as the expected table. This gives a clue to the construction of the expected table. If $\frac{a}{a+c} = \frac{b}{b+d}$, each ratio is equal to $\frac{a+b}{a+c+b+d} = \frac{a+b}{N}$. The entries in the table, $a$ and $b$, must satisfy the relations

$$a = \frac{(a+b)(a+c)}{N}, \quad b = \frac{(a+b)(b+d)}{N}$$

and it follows that

$$c = \frac{(a+c)(c+d)}{N}, \quad d = \frac{(b+d)(c+d)}{N}.$$

These numbers, then, will be the entries in the expected table and the observed and expected table will agree in all four marginal totals. We need calculate only one expected value and the rest can be found by subtraction from the marginal totals. The $\chi^2$ will therefore have 1 d.f.

If this procedure is carried out algebraically, $\sum \frac{(O-E)^2}{E}$ simplifies into $\frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ providing a simpler calculation for the $2 \times 2$ table. Tables of this sort are called <u>contingency</u> tables, and the test we have just carried out is sometimes called a test of independence of the rows and columns of the

table. Presumably this means that, in the expected table, we see the same proportion, or contrast, in one column as we do in the other. This is, of course, precisely the notion of interaction (or lack of it) coming up in this context.

We can encounter contingency tables with any number of rows and columns, $r$ and $c$ say. To test the independence of rows and columns, the expected table is calculated according to the same rules as in the $2 \times 2$ table and $\chi^2$ has $(r-1)(c-1)$ d.f.

Sampling Theory

Often populations are sampled to make absolute estimates of quantities, rather than to make comparisons, as is done in experiments. These quantities may be total amounts of something, averages or proportions. Sometimes the intention is to make predictions.

Examples:

1. Wheat (and other) farms are sampled every summer to supply a basis for predicting the total crop.

2. The labor force is sampled periodically to estimate the number and proportion of unemployed.

3. Populations of people are forever being sampled to find out their intentions in a coming election, their preferences among television programs, what kind of soap they use and so on. These samplings are called public opinion polls and market surveys.

4. Stands of timber are sampled to estimate the total stand and bodies of water are sampled to estimate the number of fish in them. These uses raise some special problems which are not generally discussed in standard texts.

Such uses of sampling are usually discussed under the heading Theory of Sampling, even though the whole of statistics is, of course, concerned with sampling. Sampling in these circumstances is, on the whole,

vastly more difficult than performing experiments, for a considerable number of reasons.

1. The population may be difficult or impossible to specify, in the sense that we can always be sure that a given observation does, in fact, belong to it.

2. The population may be changing during the course of the sampling.

3. The possibility of bias in the sampling assumes overwhelming importance, in contrast with experimental situations, where we can "design out" sources of bias and make comparisons within them.

4. Often it is necessary to use routinely gathered records and observations, which are notoriously untrustworthy.

5. Randomness can be difficult or impossible to arrange. This raises a large number of serious reservations about the extent to which any conclusions are warranted, but commonly they are ignored or glossed over.

All of these difficulties become especially acute in the sampling of populations of people, but no sampling study can be undertaken lightly. Approached properly, they are sure to be demanding and expensive.

The taking of the sample

Assuming that we are dealing with a well-defined population, the simplest kind of sample is one chosen entirely randomly. (At least, it

is the simplest sample to talk about.) Indeed, in this context, this kind of sampling is called simple random sampling. With this kind of sampling, methods already studied are sufficient to estimate the mean and the variance, to calculate confidence limits for whatever we are estimating, and so on.

Simple random sampling is rarely practised, because always we know a good deal about the structure of the population and can put this knowledge to use to improve the coverage of the sample, to diminish the error of our estimates and to make the mechanical job of getting the sample easier to manage. (Simple random sampling can be very hard indeed to carry out.) In particular, this knowledge can be used to divide the population into strata, within which the variation (which makes for error) is smaller than over the population as a whole. The object here is the same as in the blocking that is practised in designing experiments. Another purpose may also be served, because it may be desirable to sample the various strata at different intensities. In any event, having settled on the stratification to be used, we either sample each stratum randomly or we make a selection of strata (randomly) and sample each of these strata randomly.

We have now not one sample but several. We can combine the

findings from the individual samples to form our estimates for the whole population and their standard errors. Details of the procedures for carrying this out are given in books on sampling theory.

An example

As part of an elaborate study of the pheasant population of Pelee Island, an estimate of the number of nesting pheasants was required. It was known from earlier experience that the number of nests to be expected on a given area varied widely from place to place on the island, so that some kind of stratification was needed. Using knowledge gained from previous study, maps and some field work, the investigators were able to list all the regions of the island into types, ranging from highly preferred to impossible. Each observation consisted of a count of the number of nests on one acre. The list of types then read as follows.

Type A, containing a acres, preference 1
B, b , 2
.
.
.
E e , impossible.

Presumably, then, the variation within types would be considerably less

than variation between types and the types furnish a sensible basis for stratification.

Now, it seems obvious that type A ought to be sampled most heavily, type B somewhat less, ... type E not at all. The following considerations bear on the choice of the sampling rates. Let us say that we will choose randomly $n_1$ of the $a$ sampling units in type A, $n_2$ of the $b$ units in type B, and so on and let us call $x_1$, $x_2$, ... the total counts in type A, type B, etc. ($x_1$ is the sum of $n_1$ counts, etc.) The estimate of the total number of nests on the island will be

$$T = x_1 \frac{a}{n_1} + x_2 \frac{b}{n_2} + \ldots$$

$$= a\bar{x}_1 + b\bar{x}_2 + \ldots \quad .$$

We would, let us say, like the variance of $T$ to be as small as possible, for a given total number $(N)$ of acres sampled. If $\sigma_1^2$, $\sigma_2^2$, ... represent the variances within types we have

$$\text{Var } T = a^2 \text{ Var } \bar{x}_1 + b^2 \text{ Var } \bar{x}_2 + \ldots$$

$$= a^2 \frac{\sigma_1^2}{n_1} + b^2 \frac{\sigma_2^2}{n_2} + \ldots \quad .$$

This expression we require to be as small as possible, subject to

the condition that $n_1 + n_2 + \ldots = N$ . Thus, we write

$$\psi = \frac{a^2\sigma_1^2}{n_1} + \frac{b^2\sigma_2^2}{n_2} + \ldots + \lambda(n_1 + n_2 + \ldots - N)$$

$$\frac{\partial\psi}{\partial n_1} = -\frac{a^2\sigma_1^2}{n_1^2} + \lambda = 0 \ ,$$

$$\frac{\partial\psi}{\partial n_2} = -\frac{b^2\sigma_2^2}{n_2^2} + \lambda = 0 \ ,$$

etc.

Therefore, each $n_i$ is to be chosen proportional to the product of the number of units (acres) within the type and the standard deviation within the type. Of course, the standard deviations would not be known (they can be estimated after the sampling is completed), but experience with this sort of thing suggests that they are likely to be roughly proportional to the average counts within the types which, in turn, might be guessed with fair accuracy on the basis of prior experience.

Even though there is some uncertainty about the best sampling rates to use, there is every reason to expect a stratified random sampling of this sort to be vastly superior to simple random sampling. The effort is concentrated where it will do the most

good and the error the estimate is exposed to (the within stratum error) is likely to be much smaller than that reflecting variation over all the strata.

This example is simpler than most. The population is well defined and, for most part, easily accessible (there was some difficulty with brambles). Much dependable prior knowledge is available. There are, it seems, none of the sources of bias one fears in many sampling studies.

Some non-parametric tests of significance.

The sign test.

One way of approaching the paired comparison type of experiment is to list the differences between pairs and treat them as observations from a single population. The test of significance for the difference between two samples becomes a test for significant departure from zero of the average difference.

If, instead of the numerical values of the differences, we list simply their signs, the supposition that the two samples come from identical populations implies that positive and negative differences are equally probable and that the set of signs could be considered to come from a binomial population with $p = \frac{1}{2}$ . (Strictly speaking,

this argument holds only for continuous variables.) This remark forms the basis of a test based on the binomial distribution.

Because measurements are always discrete, it may happen that some of the differences have value zero (the problem of ties.) Such differences may be dropped and the sample size reduced accordingly.

It seems obvious that the sign test is less discerning (powerful) than the t - test, because it ignores the numerical values of the differences. On the other hand, it requires little calculation and may, on occasion, be all that is required. Also, there are situations in which, for one reason or another, the numbers needed to calculate a t - test are not available. See exercise 31.

The Mann-Whitney Test.

When the observations are obtained in a completely randomized fashion, the sign test cannot be used. Suppose we have two samples, $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$, taken in circumstances which admit that the populations may be different, in particular, that they have different locations. We require a test of the supposition that the populations are identical.

We would presumably have little doubt that the populations are different if, for example, each $x$ is less than every $y$, because this outcome is highly improbable if the populations are identical. In less extreme cases, we require some measure to display the preponderance of cases in which $x < y$.

One way of defining such a measure is to order each sample in increasing order of magnitude, getting say $[x_1]$, $[x_2]$, ... $[x_{n_1}]$ and $[y_1]$, $[y_2]$, ... $[y_{n_2}]$, then pool the samples and order the whole set in increasing order. The pooled and ordered samples might then be, for example,

$$[x_1] \quad [x_2] \quad [y_1] \quad [x_3] \quad [x_4] \quad [x_5] \quad [y_2] \quad \ldots \quad [y_{n_2}]$$

Now, calculate numbers as follows:

$u_1$ = number of x's less than $[y_1]$ (for the example $u_1 = 2$)

$u_2$ = number of x's less than $[y_2]$ ($u_2 = 5$)

.

.

.

$u_{n_2}$ = number of x's less than $[y_{n_2}]$ ($u_{n_2} = n_1$)

$$U = \sum_{i=1}^{n_2} u_i \ .$$

One extreme case, when each $x$ is less than every $y$, yields $u_1 = u_2 = \ldots = u_{n_2} = n_1$ and $U = n_1 n_2$. The other extreme, in which each $x$ is greater than every $y$, evidently yields $U = 0$. Other arrangements yield values between these two.

Under the supposition that the two samples came randomly from identical populations, all orders of the $n_1 + n_2$ observations are equally probable and the probability of obtaining each possible value of $U$ can be calculated by direct combinatorial methods. We can then select those values so extreme as to be highly improbable, if the samples did in fact come from identical populations and therefore, if one of them is obtained, it can reasonably be concluded that the samples must have come from different populations.

For small values of $n_1$ and $n_2$, the probability distribution of $U$ is tabulated. For large $n_1$ or $n_2$, $U$ is approximately normal, with mean $\frac{n_1 n_2}{2}$ and variance $\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$.

## Appendix

### Orthogonal Transformations

A linear transformation which changes a set of values or variables, $x_1, x_2, \ldots x_n$, to another set $y_1, y_2, \ldots y_n$, may be written in the form

$$y_1 = a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

$$y_2 = b_1y_1 + b_2x_2 + \ldots + b_nx_n$$

.

. and so on.

.

$$y_n =$$

The symbols $a_i, b_i \ldots$ stand for numerical constants. If these constants satisfy relations like $a_1^2 + a_2^2 + \ldots a_n^2 = 1$, $b_1^2 + b_2^2 + \ldots + b_n^2 = 1$, $a_1b_1 + a_2b_2 + \ldots + a_nb_n = 0$, for every pair of y's, the transformation is said to be orthogonal and normalized (orthonormal). It will be called henceforth simply orthogonal.

Orthogonal transformations are useful in statistics for extracting, from a set of observations, what they have to say about the various sources of variation which gave rise to them. In order that this shall be possible, the observations must be made in certain patterns which are themselves said to be orthogonal.

The reason why orthogonal transformations are used, rather than some others, lies in some facts which can be established

mathematically.

(1) If the x's can be regarded as independent, normal variables with variance $\sigma^2$ , the y's may also be so regarded.

(2) $y_1^2 + y_2^2 + \ldots y_n^2 = x_1^2 + x_2^2 + \ldots x_n^2$

The writing of orthogonal transformations.

When numbers are used instead of symbols to construct a specific transformation, of necessity the numbers are fractions or square roots of fractions. It is convenient, to avoid an inordinate amount of writing, to use integers, chosen to satisfy the condition that the sums of products must be zero and determine the divisors so that the squares of the coefficients add to one. To illustrate the construction of any orthogonal transformation, take $n = 3$ . To write

$$y_1 = a_1x_1 + a_2x_2 + a_3x_3$$

we may choose any three numbers, 1, 2 and 3 , say, and divide them by the square root of the sum of their squares. Then,

$$y_1 = \frac{1}{\sqrt{14}} x_1 + \frac{2}{\sqrt{14}} x_2 + \frac{3}{\sqrt{14}} x_3$$

To write $y_2 = b_1x_1 + b_2x_2 + b_3x_3$ , orthogonal to $y_1$ , we must choose $b_1, b_2, b_3$ so that

$$b_1 + 2b_2 + 3b_3 = 0 .$$

Any set of b's satisfying this equation will serve. It is simplest to choose integers, e.g. $b_1 = -7$ , $b_2 = 2$ , $b_3 = 1$ . Then

$$y_2 = \frac{-7}{\sqrt{54}} x_1 + \frac{2}{\sqrt{54}} x_2 + \frac{1}{\sqrt{54}} x_3 .$$

To obtain $y_3 = c_1x_1 + c_2x_2 + c_3x_3$ , orthogonal to $y_1$ and $y_2$ values of $c_1$ , $c_2$ , $c_3$ must be found such that

$$c_1 + 2c_2 + 3c_3 = 0 .$$

$$-7c_1 + 2c_2 + c_3 = 0 .$$

Integers satisfying both equations are $c_1 = 2$ , $c_2 = 11$ , $c_3 = -8$ .

$$y_3 = \frac{2}{\sqrt{189}} x_1 + \frac{11}{\sqrt{189}} x_2 - \frac{8}{\sqrt{189}} x_3 .$$

In writing $y_1$ , there is a free choice of two numbers, the third being determined by the fact that the sum of their squares must be unity. In writing $y_2$ , only one choice may be made and in writing $y_3$ , there is no choice whatever.

The transformation can be exhibited most simply in the form

| | $x_1$ | $x_2$ | $x_3$ | divisor |
|---|---|---|---|---|
| $y_1$ | 1 | 2 | 3 | $\sqrt{14}$ |
| $y_2$ | -7 | 2 | 1 | $\sqrt{54}$ |
| $y_3$ | 2 | 11 | -8 | $\sqrt{189}$ |

In statistical applications, an orthogonal transformation is constructed to exhibit the contributions of all the sources of variation provided for in obtaining the observations. The first component, $y_1$ , in the applications studied here, is given a set of equal coefficients, i.e. $a_1 = a_2 = \ldots = a_n$ . This choice decrees that $y_2$ , $y_3$ , $\cdots$ $y_n$ must have coefficients that sum to zero. For this reason, they are called <u>contrasts</u> or comparisons or differences.

Experiments are wholly concerned with contrasts, with changes

brought about in a response variable by changing causal variables. The way in which the transformation is constructed is dictated by the nature of the changes introduced with the causal system.

The orthogonal transformation is used as an instrument for understanding and exhibiting the structure of an experiment and the contrasts it is intended to study. Ordinarily the transformation is not set up as a device for making calculations, but good calculating rules are derived from it. Usually it is sufficient to write the transformation for a diminutive example of an actual experiment.

In complicated experiments, the transformation is useful in displaying, for each component, which sources of variation are included in it and which ones are excluded. When the coefficients of a component sum to zero within each level of a source of variation, variation arising from that source is excluded from the component; when the coefficients do not sum to zero within each level, the source of variation does contribute to the value of the component. When the coefficients of a component sum to zero within each level of every source of systematic variation provided in the arrangement of the experiment, the component gathers up only variation of the sort attributed to error.

Crossed Classifications.

When observations are classified according to two or more criteria where each level of each criterion occurs in combination

with each level of every other criterion, the classification may be said to be crossed. This occurs, for example, in factorial experiments and randomized blocks. In these circumstances, the possibility exists of perceiving the effects of interactions. This discussion has to do with the writing of interaction components in the orthogonal transformation.

To discuss an example, think of a $3 \times 3$ factorial arrangement, with one factor $A$ at levels $a_1$, $a_2$, $a_3$ and another factor $B$ at levels $b_1$, $b_2$, $b_3$. Let $x_{ij}$ stand for an observation, or better, the average of several observations, arising out of an orthogonal experiment, on the combination $a_i b_j$. Then, apart from errors, we can think of the structure of $x_{ij}$ as $x_{ij} = \mu + \alpha_i + \beta_j + \tau_{ij}$, which says simply that our observations may be expected to vary systematically from one level of $A$ to another, from one level of $B$ to another and, in addition, in a way which cannot be accounted for in either of these ways (this is the interaction). It is convenient, in this model, to make $\Sigma\alpha_i = 0$, $\Sigma\beta_j = 0$, $\sum_i \tau_{ij} = 0$ for all $j$ and $\sum_j \tau_{ij} = 0$ for all $i$. This convention is not needed here, however.

If we start writing a transformation of the $x_{ij}$ to display, first, main effects, we may choose three numbers, $c_1$, $c_2$, $c_3$, such that $c_1 + c_2 + c_3 = 0$, and use them to display an $A$ main effect as shown.

| | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{31}$ | $x_{32}$ | $x_{33}$ |
|---|---|---|---|---|---|---|---|---|---|
| $y_a$ | $c_1$ | $c_1$ | $c_1$ | $c_2$ | $c_2$ | $c_2$ | $c_3$ | $c_3$ | $c_3$ |
| $y_b$ | $d_1$ | $d_2$ | $d_3$ | $d_1$ | $d_2$ | $d_3$ | $d_1$ | $d_2$ | $d_3$ |
| $y_a x y_b$ | $c_1d_1$ | $c_1d_2$ | $c_1d_3$ | $c_2d_1$ | $c_2d_2$ | $c_2d_3$ | $c_3d_1$ | $c_3d_2$ | $c_3d_3$ |

similarly, a B main effect may be displayed by choosing any three numbers $d_1$ $d_2$ $d_3$ with $d_1 + d_2 + d_3 = 0$ . Now, it is asserted that an interaction component, corresponding to $y_a$ and $y_b$ , can be formed simply by multiplying coefficients, column by column, of $y_a$ and $y_b$ .

The component labelled $y_a \times y_b$ can be written out as

$$y_a \times y_b = \sum_i \sum_j c_i\, d_j\, x_{ij} = \sum_i \sum_j c_i\, d_j(\mu + \alpha_i + \beta_j + \tau_{ij}) \ ,$$

which can be expanded and rearranged to read

$$\mu \sum_i c_i \sum_j d_j + \sum_i c_i\, \alpha_i \sum_j d_j + \sum_j d_j\, \beta_j \sum_i c_i + \sum_i \sum_j c_i\, d_j\, \tau_{ij} \ ,$$

which reduces to $\sum_i \sum_j c_i\, d_j\, \tau_{ij}$ , because $\sum c_i = \sum d_j = 0$ .

Calculations of the same kind yield:

$$y_a = \sum_i \sum_j c_i\, x_{ij} = \sum_i \sum_j c_i(\alpha_i + \tau_{ij})$$

$$y_b = \sum_i \sum_j d_j\, x_{ij} = \sum_i \sum_j d_j(\beta_j + \tau_{ij}) \ .$$

From these expressions we see that contributions from the $\tau_{ij}$ are perceived in the component $y_a \times y_b$ and that, if they turn out to be different from zero, their values distort the main effect components $y_a$ and $y_b$ , rendering them useless.

N.B. It has not been asserted that interaction components can be constructed only by the column by column multiplication rule. There are, in fact, other ways, leading to different sets of interaction components.

Exercises

1. The weight of coating, in hundredths of an ounce per square foot, is measured for 60 sheets.

| | | |
|---|---|---|
| 147 | 160 | 158 |
| 162 | 160 | 132 |
| 152 | 138 | 158 |
| 177 | 173 | 162 |
| 155 | 170 | 147 |
| 153 | 160 | 142 |
| 138 | 160 | 145 |
| 137 | 151 | 134 |
| 164 | 153 | 144 |
| 146 | 159 | 158 |
| 163 | 150 | 148 |
| 154 | 160 | 148 |
| 153 | 154 | 155 |
| 134 | 159 | 164 |
| 160 | 149 | 157 |
| 137 | 155 | 162 |
| 146 | 158 | 148 |
| 166 | 148 | 175 |
| 132 | 160 | 150 |
| 165 | 148 | 172 |

(a) Make a grouped frequency distribution, using about 10 classes.

(b) Plot the histogram of the distribution.

(c) Plot the cumulative frequency distribution.

(d) Obtain the median of the distribution.

(e) Calculate the average of the frequency distribution.

(f) Calculate the second moment about the average of the frequency distribution.

(g) Compare the numbers in (e) and (f) with the average and second moment about the average, calculated directly from the sample. The sum and sum of squares of the sample values are, dropping the first digit, 3223 and 179797.

2. An urn contains three red, four white and five blue balls. Three balls are to be drawn randomly without replacement. Let R stand for the event: at least one red ball will be drawn; r - exactly one red ball will be drawn, with similar definitions for W, w, B, b .

(a) Calculate the probabilities of the events r, R, b, B, w, W, rW, WBR, R + W, RW .

(b) Calculate P(B/RW) and P(r/W) .

3. A deck of six cards, three 7's and three 8's, is to be dealt to three players, A, B and C . List the various different distributions of hands among the three players (there are seven of them). Under the assumption that the cards are to be distributed randomly to the players, calculate the probability of each of the possible distributions. Now calculate the probabilities of the following events.

(a) A gets a pair.

(b) One of the players gets a pair.

(c) Exactly one of the players gets a pair.

(d) B gets an 8 , given that A gets an 8 .

(e) B gets an 8 , given that A gets two 8's .

(f) B gets an 8 , given that A gets exactly one 8 .

4. A lottery in which N tickets are sold provides K prizes to be awarded by selecting randomly K of the N tickets. A gambler buys two tickets. Calculate in two ways the probability that he will be a winner.

5. An urn contains a white balls and b black balls. One ball is to be selected randomly and removed. A second ball is then to be drawn.

(a) Calculate the probability that the second ball will be white.

(b) If the second ball proves to be white, what is the probability that the first one was white?.

6. A coin is to be thrown until head has turned up twice. Find the probability distribution of the number of throws.

7. Two defective light bulbs have been mixed with two good ones. A bulb is to be selected randomly from the four bulbs and tested. If it proves to be defective it will be destroyed and the process of selecting and testing will be continued. Let x represent the

number of bulbs that will be tested until a good one is found. Determine the probability distribution of x . Calculate the mean and variance of this distribution.

8. An infinite population is specified by the following frequency distribution.

| x | $p_x$ |
|---|---|
| 1 | 1/4 |
| 2 | 1/2 |
| 3 | 1/4 |

(a) Calculate the mean and the variance of this distribution.

(b) List all possible samples of 2 observations that can be drawn from this population.

(c) Calculate the probability of each sample, under the supposition that it is drawn randomly.

(d) Calculate the average of each sample and form the probability distribution of these averages.

(e) Calculate the mean and the variance of the distribution in (d).

(f) How can the mean and the variance found in (e) be calculated without forming the distribution in (d)?

(g) Same as above, using samples of 3 observations.

9. Large consignments are accepted or rejected on the basis of a

randomly chosen sample of 20 items. If the sample contains 3 or more defective items, the consignment is rejected. What is the probability that it will be rejected if it contains (a) 1% ; (b) 10% ; (c) 30% defective items?

10. On a true-false test, a candidate answered correctly

(a) 7 out of a total of 10 questions;

(b) 70 out of a total of 100 questions.

In each instance, decide whether one can reasonably maintain that all the answers were simply guesses.

11. An unbiased coin is to be tossed 15 times. Calculate the probability that

(a) the number of heads will be outside the range 4 to 10;

(b) neither the number of heads nor the number of tails will be outside the range 4 to 10.

12. Let $x$ be a binomial variable with $n = 25$ and $p = 0.30$. Calculate $P(5 \leq x \leq 13)$ using binomial tables and using the normal approximation to the binomial distribution, both with and without the correction for continuity.

13. A sample of 25 observations has been drawn from a population asserted to be normal with mean 6 and variance 4.

(a) The average of the sample is 8. Show that this result is in serious conflict with the assertion.

(b) The sum of the squares of the observations is 3136. Calculate test functions that bear on the following questions.

(i) Is the specified variance incorrect?

(ii) Is the specified mean incorrect?

(c) Calculate 95% confidence limits for the mean of the population.

14. A sample of 25 observations, drawn randomly from a normal population, yields an average 43.7 feet.

(a) Calculate 99% confidence limits for the mean of the population, given that the variance of the population is 9 square feet.

(b) If the variance of the population is not known, but is estimated from the sample by calculating $s^2 = 10.4$ , how is the procedure for setting confidence limits modified from that used in part (a)? Calculate 99% confidence limits in this instance.

(c) How are confidence limits, like those calculated in (a) and (b), to be interpreted?

15. An established method for determining the specific gravity of metals is known, through extensive use, to have normally distributed errors with standard deviation $\sigma = 0.10$ . Two students, A and B , use the method to measure the specific gravity of a given piece of metal, with the following results.

A : 10.20, 10.10, 10.30, 10.50, 10.40, 10.10, 10.50 .

B : 10.30, 10.28. 10.32, 10.29, 10.31 .

Do you think that either A or B is using the method properly?

16. Two methods of teaching reading are compared by dividing a group of 10 children randomly into two groups, one taught by method A the other by method B . At the end of the trial, each child is required to read the same passage. The numbers of mistakes are recorded below.

| method A | method B |
|---|---|
| 39 | 32 |
| 47 | 41 |
| 51 | 30 |
| 32 | 37 |
| 43 | 34 |

Do these results furnish acceptable evidence of a genuine difference between the methods?

If you had the responsibility of designing an experiment to compare two methods of teaching reading, with the intention of selecting one of them to be used throughout Ontario, what would you recommend?

17. Two types of coating, A and B , intended to retard corrosion of iron pipes, are compared by putting A on one half of each of a number of specimens of pipe and B on the other half. They are then buried in soil for a year, removed and the depths of pits caused by corrosion are measured. The depths of the deepest pits are given in the following table.

| specimen no. | coating A | coating B | untreated |
|---|---|---|---|
| 1 | 51 | 73 | 81 |
| 2 | 41 | 43 | 52 |
| 3 | 43 | 47 | 55 |
| 4 | 41 | 53 | 63 |
| 5 | 47 | 58 | 65 |
| 6 | 32 | 47 | 50 |
| 7 | 24 | 53 | 62 |
| 8 | 43 | 38 | 48 |
| 9 | 53 | 61 | 58 |
| 10 | 52 | 56 | 59 |

Decide, on the basis of a test of significance, whether these results indicate a difference between the coatings. Carry out the arithmetical calculations in two ways, the analysis of variance and the paired-comparison.

Suppose now that the experiment had been carried out somewhat differently, to provide information on whether the coatings are any use at all. To this end, each specimen is marked off in three equal segments, one to receive A , one B and the third left uncoated. Presumably the allocation of treatments to segments would be made randomly for each specimen.

Calculate the analysis of variance table for this experiment and make such tests of significance as you think are needed.

18. Construct a set of $2 \times 2$ tables in the following way.

(a) Choose any number and enter it in each cell of the table.

(b) Choose any two numbers which add to zero, add one of them to each number in row 1 and the other to each number in row 2.

(c) Choose any two numbers which add to zero, add one of them to each number in column 1 of table (b) and the other to each number in column 2.

(d) Choose any four numbers which have the property that, when arranged in a square, the sum in each row and each column is zero. Assign these numbers to the cells of table (c) and add them to the numbers already there.

Calculate an analysis of variance table for each of the four tables you have constructed, separating the variation into that attributable to rows, columns and rows × columns.

Question

If the numbers added to the rows in step (b) do not add to zero, where, in your analysis, would this fact show up?

19. The effects of three drugs on level of performance and learning (as displayed by changes in performance over successive trials) are compared using 15 subjects, 5 allocated randomly to each drug.

| drugs | subjects | trials | | |
|---|---|---|---|---|
| | | $A_1$ | $A_2$ | $A_3$ |
| | $B_1$ | 2 | 4 | 7 |
| | $B_2$ | 2 | 6 | 10 |
| $D_1$ | $B_3$ | 3 | 7 | 10 |
| | $B_4$ | 7 | 9 | 11 |
| | $B_5$ | 6 | 9 | 12 |
| | $B_6$ | 5 | 6 | 10 |
| | $B_7$ | 4 | 5 | 10 |
| $D_2$ | $B_8$ | 7 | 8 | 11 |
| | $B_9$ | 8 | 9 | 11 |
| | $B_{10}$ | 11 | 12 | 13 |
| | $B_{11}$ | 3 | 4 | 7 |
| $D_3$ | $B_{12}$ | 3 | 6 | 9 |
| | $B_{13}$ | 4 | 7 | 9 |
| | $B_{14}$ | 8 | 8 | 10 |
| | $B_{15}$ | 7 | 10 | 10 |

(a) Plot such graphs as you think may be useful.

(b) What kind of experimental plan is this?

(c) Calculate the analysis of variance and make any tests of significance you require to reach conclusions about the effects of the drugs. Do these conclusions agree with what you perceive in the graphs.

20. A $3^2$ factorial experiment, carried out in a randomized block design with two replications, yields the observations given below. Calculate an appropriate analysis of variance and make the tests of significance that are required. (Assume that the levels of both a and b are equally spaced.) Plot the graphs indicated by the analysis.

| | Rep. 1 | | | Rep. 2 | | |
|---|---|---|---|---|---|---|
| | $a_1$ | $a_2$ | $a_3$ | $a_1$ | $a_2$ | $a_3$ |
| $b_1$ | 19.86 | 26.37 | 29.72 | 20.88 | 24.38 | 29.64 |
| $b_2$ | 15.35 | 22.82 | 27.12 | 15.86 | 20.98 | 24.27 |
| $b_3$ | 4.01 | 10.34 | 15.64 | 4.48 | 9.38 | 14.03 |

The "preliminary" analysis of variance table is

| | d.f. | s.s. |
|---|---|---|
| replications | 1 | 2.9850 |
| treatments | 8 | 1077.2496 |
| error | 8 | 7.2699 |
| total | 17 | 1087.5045 |

21. Think of a chemical process whose yield of a certain chemical is measured at five equally-spaced temperatures, which may as well be labelled 0, 1, 2, 3, 4 . Suppose the measured yields are,

in appropriate units:

| | | | | | |
|---|---|---|---|---|---|
| (a) | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |
| (b) | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
| (c) | 1.1 | 1.3 | 1.6 | 2.0 | 2.5 |
| (d) | 1.1 | 1.4 | 2.0 | 3.0 | 4.5 |
| (e) | 1.1 | 1.5 | 2.5 | 4.5 | 7.0 |

Plot the curve of yield against temperature for each of the sets (a), (b), (c), (d), (e) .

Apply to each of the sets the transformation whose coefficients are listed below.

| | | | | | |
|---|---|---|---|---|---|
| average | 1 | 1 | 1 | 1 | 1 |
| linear | -2 | -1 | 0 | 1 | 2 |
| quadratic | 2 | -1 | -2 | -1 | 2 |
| cubic | -1 | 2 | 0 | -2 | 1 |
| quartic | 1 | -4 | 6 | -4 | 1 |

How would you use this transformation, in an actual experimental situation, to decide on the nature of the curve required to describe the relation between yield and temperature? Suppose the yield is determined three times at each temperature, that the error s.s. , with 10 d.f. , is 11.3 and that the <u>average</u> yields are found to be

1, 2, 9, 28, 64.

Plot the graph and carry out an analysis to ascertain the degree of the polynomial that fits the points adequately.

22. A factorial experiment, in which two levels of a factor A are tested in all combinations with two levels of a factor B, leads to an analysis of variance table

| | d.f. | (a) s.s. | (b) s.s. | (c) s.s. |
|---|---|---|---|---|
| A | 1 | 96 | 7 | 96 |
| B | 1 | 7 | 5 | 96 |
| A×B | 1 | 5 | 96 | 96 |
| error | 8 | 48 | 48 | 48 |

Carry out, in each of the situations (a), (b), (c), the tests of significance you think are required and state the conclusions they indicate.

23. Two kinds of sole leather, A and B, are to be compared by issuing boots made with each kind of leather to a squad of soldiers who are going on a route march. At the conclusion of the march, the amount of wear of each boot will be measured in some suitable way.

Two plans for carrying out the test have been put forward.

Plan 1. The squad will be divided into two equal groups, each with n soldiers in it. Boots made with leather A will be issued to one group and boots made with leather B to the other group.

Plan 2. Each of the $2n$ soldiers will be issued a pair of boots, one made with leather $A$, the other with leather $B$.

Compare and contrast the two plans. Where, in each of them, is randomness needed? Set up an analysis of variance table for each plan and indicate the proper error term for testing the $A$ vs. $B$ comparison. Which of the two plans would you expect to furnish the more precise comparison?

24. A trial, carried out according to Plan 2, yields the following measurements of wear.

| individual | leather A | leather B |
|---|---|---|
| 1 | 5 | 7 |
| 2 | 3 | 4 |
| 3 | 3 | 2 |
| 4 | 4 | 5 |
| 5 | 2 | 4 |
| 6 | 2 | 3 |
| 7 | 5 | 6 |
| 8 | 3 | 6 |
| 9 | 4 | 3 |
| 10 | 6 | 7 |

Carry out an analysis of these results, leading to a test of significance of the average difference between $A$ and $B$.

25. Let $y$ stand for the height in inches of a man and $x$ for the height of his father. Fit a regression equation $Y = \bar{y} + b_1(x - \bar{x})$

to the set of (x,y) pairs.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| y : | 66 | 65 | 67 | 68 | 69 | 69 | 71 | 69 |
| x : | 60 | 62 | 64 | 66 | 68 | 70 | 72 | 74 |

(a) Calculate the s.s. attributable to regression and the s.s. of residuals. Enter them in an analysis of variance table.

(b) Is $b_1$ significantly different from zero?

(c) Is $b_1$ significantly different from one?

(d) Use the regression equation to estimate the average height of sons whose fathers are 76 inches tall.

(e) Plot the equation $Y = \bar{y} + b_1(x - \bar{x})$ on a graph.

(f) Calculate 95% confidence limits for $E(y/x)$ at several selected values of $x$. Plot these points on the graph and use them to sketch in the boundaries of the confidence band.

26. Go back to the observations listed in exercise 21, relating yield with temperature. Fit a regression $Y = \bar{y} + b_1(x - \bar{x})$ to these observations (y representing yield, x temperature). Calculate the s.s. attributable to regression and the s.s. of residuals. Test the s.s. of residuals against error to decide if the regression equation fits well enough. (Of course, the answer to this question is known from 21.)

27. An experiment is carried out to study the dependence of the stiffness, y, of fabric on three factors : D, the diameter of the weft yarn, T, the amount of twist in the weft yarn and x,

the number of weft yarns to the inch.

For obvious reasons, the factor $x$ could not be introduced, over a sufficiently wide range, orthogonally with the factors $D$ and $T$. The layout of the experiment and the measurement of stiffness are given in the following table.

| | $D_1 = 2.1$ | | $D_2 = 3.0$ | |
|---|---|---|---|---|
| | $y$ | $x$ | $y$ | $x$ |
| $T_1 = 2$ | 84 | 88 | 104 | 95 |
| | 75 | 82 | 94 | 88 |
| | 61 | 74 | 88 | 84 |
| | 71 | 95 | 78 | 95 |
| $T_2 = 8$ | 66 | 92 | 72 | 92 |
| | 54 | 84 | 56 | 82 |

(a) How would you describe the structure of this experiment?

(b) Plot accurately on the same graph the $(x,y)$ points for each of the sets $D_1T_1$, $D_1T_2$, $D_2T_1$, $D_2T_2$.

By inspection of these graphs, answer the following questions.

(c) How would you describe the dependence of $y$ on the factor $x$? How might you specify it numerically?

(d) Does the factor $x$ interact with one or both of the factors $D$ and $T$?

(e) Do the factors $D$ and $T$ interact?

(f) Read, from the four lines of the graph, the y-values corresponding to $x = 88$, which is approximately the average of all the x-values in the experiment. Use these four y-values to calculate the change in stiffness brought about by changing D from $D_1$ to $D_2$ and by changing T from $T_1$ to $T_2$. These changes may be said to be adjusted, inasmuch as they do not depend on x.

The numerical accompaniment to this graphical analysis is called the analysis of covariance.

28.(a) A count of the number of thunderstorms occuring over a large area in one year yields the following list.

| | |
|---|---|
| June | 60 |
| July | 100 |
| August | 80 |

Are these observations inconsistent with the supposition that thunderstorms are distributed equally among the three months?

(b) A similar count made the following year yields the list

| | |
|---|---|
| June | 80 |
| July | 100 |
| August | 60 |

Do the two lists indicate a genuine difference, between one year and another, in the distribution of thunderstorms over months?

29. An angler recorded his catch for a season in the following form.

| | morning | midday | evening |
|---|---|---|---|
| bass | 60 | 20 | 40 |
| pike | 70 | 50 | 30 |
| perch | 90 | 100 | 60 |

Do these records indicate differences, between one kind of fish and another, in the pattern of success throughout the day?

30. Go back to exercise 17 and carry out a non-parametric test of the difference between the two coatings.

31. A department of education carried out a trial in 12 schools, to compare two methods of teaching reading. It reported that, in every school, method A scored better than method B ; but that the overall difference between the methods was not significant. Can you accept this statement?

32. A water diviner advertised that he had been consistently successful in locating water in 90% of his attempts. The following year he made 100 attempts and located water in 70 of them.
Do you think his skill had diminished? What suppositions do you make in reaching your conclusion.

33. The experience of airlines shows that 10% of ticket holders do not show up for their flights. If an airline overbooks by 5%, what is the probability that a flight, so overbooked, that can accommodate 100 passengers, will not be able to accommodate all the passengers who will arrive for it?
Write an exact expression for this probability and approximate it as well as you can.

34. A process produces ball bearings with mean weight 1 ounce and standard deviation .001 ounce. The ball bearings are packaged in boxes of 100.
(a) Calculate the mean and the standard deviation of the weights of the boxes. (Ignore the weights of the containers.)
(b) Nine boxes of ball bearings are purchased. Their weights prove to be, in ounces.

99.70 , 99.71 , 99.72 , 99.69 , 99.68 , 99.70

99.69 , 99.71 , 99.70 .

Do you think that these 9 boxes can be regarded as a randomly chosen sample from the population described in (a) ?

35. Grade 1 wheat has mean weight 65.0 pounds per bushel, with standard deviation 0.2 pounds, 5 bushels of wheat are chosen randomly from a large consignment not known to be Grade 1. Their

weights prove to be, in pounds,

64.4 , 65.3 , 64.6 , 65.9 , 64.8 .

(a) Is the consignment more variable than Grade 1 wheat?

(b) Calculate 95% confidence limits for the mean weight per bushel of the consignment.

36. Three specimens of each of five different metals were immersed in a corrosive solution and the rate of corrosion of each specimen was determined

| Metal | Corrosion Rate |
|---|---|
| Aluminum | 0.5 , 0.4 , 0.6 . |
| Stainless steel | 0.6 , 0.7 , 0.6 . |
| Carbon steel | 6.5 , 7.0 , 7.3 . |
| Enamel-coated steel | 0.8 , 0.6 , 0.8 . |
| Nickel alloy | 4.1 , 3.5 , 3.0 . |

(a) What type of experimental arrangement is this?

(b) Calculate an analysis of variance and assemble the results in an analysis of variance table.

(c) Can you reasonably account for the differences among the metals on the basis of error only?

(d) Estimate the difference between the corrosion rates of stainless steel and carbon steel and determine a 90% confidence interval for this difference.

37. Three brands of detergent A, B, C, are compared by washing uniformly soiled specimens of cloth in them at two different temperatures, $T_1$ and $T_2$ . Six identical washing machines are used, so that the six combinations of detergents and temperatures may be tested simultaneously. Two replications of the test are carried out on two successive days.

After the specimens are washed and ironed, measurements of brightness are made on them. The following tables record the averages of the brightness determinations obtained under the various sets of conditions.

| | Day 1 | | | | Day 2 | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | | A | B | C |
| $T_1$ | 10.6 | 11.5 | 12.2 | $T_1$ | 11.6 | 12.4 | 13.6 |
| $T_2$ | 11.7 | 10.8 | 10.1 | $T_2$ | 12.4 | 11.9 | 11.2 |

Carry out an analysis of variance on these results and use it to reach a conclusion about the differences among the detergents. Make a table of such averages as you think are warranted.

38. The yield of a chemical process is measured, on two occasions, at each of three temperatures, $125^{\circ}$ , $150^{\circ}$ , $175^{\circ}$ . The yields, in tons, are given in the following table.

| $125^{\circ}$ | $150^{\circ}$ | $175^{\circ}$ |
|---|---|---|
| 173.6 | 192.5 | 181.7 |
| 182.4 | 198.3 | 175.2 |

Analyse these records with a view to studying the dependence of yield on temperature.

39. Two strains of virus are to be compared by applying them to leaves of plants and counting the number of lesions produced. 8 young plants, each bearing two leaves, are to be used, one strain to be applied to one leaf of each plant, the other strain to the other leaf.

(a) What kind of experimental arrangement is this?

(b) Where is randomness required?

(c) Which of the terms in the following list can properly be used in describing this experimental arrangement?
replication, block, factorial, split-plot, confounded, paired.

d) The counts yielded by the trial are given below. Carry out a test of significance on the average difference between strains.

| Plant No. | Strain A | Strain B |
|---|---|---|
| 1 | 4 | 6 |
| 2 | 7 | 8 |
| 3 | 3 | 2 |
| 4 | 6 | 9 |
| 5 | 6 | 7 |
| 6 | 8 | 10 |
| 7 | 5 | 6 |
| 8 | 2 | 3 |

40. Appetites of rats, as measured by weight of food consumed, are to be compared under the influence of a drug, to be administered at three levels, 0.1 , 0.3 , 0.5 mg. per gram of body weight, and at two different durations of starvation, 5 and 9 hours. To start with, six rats are assigned randomly to the six treatment combinations. The following day, six more rats are similarly tested, providing a second replication. The results, expressed in grams of food consumed, are given in the following tables:

| | | rep. 1 | | | | rep. 2 | | |
|---|---|---|---|---|---|---|---|---|
| | | level of drug | | | | level of drug | | |
| | | 0.1 | 0.3 | 0.5 | | 0.1 | 0.3 | 0.5 |
| duration | 5 hr. | 9.16 | 11.57 | 5.22 | 5 hr. | 11.82 | 11.53 | 9.21 |
| | 9 hr. | 16.08 | 10.30 | 9.27 | 9 hr. | 14.65 | 14.46 | 6.10 |

1. How would you describe the structure of this experiment?
2. Draw up a suitable analysis of variance table (or a sequence of them), listing sources of variation and the degrees of freedom associated with them, to correspond to this structure.
3. Calculate the requisite sums of squares and use them to decide how much reduction of the observations is warranted.
4. Discuss the nature of the graphs that may be plotted to display your conclusions.

41. A firm is introducing a new product. It wishes to study the effect on sales of using two different counter displays, $A_1$ and $A_2$, and two different prices, $B_1$(\$4), and $B_2$(\$8). 12 stores are available for the trial, 3 assigned by random allocation to each of the four display-price combinations. The unit sales recorded for the trial period are as follows.

| $A_1B_1$ | $A_1B_2$ | $A_2B_1$ | $A_2B_2$ |
|---|---|---|---|
| 64 | 72 | 60 | 88 |
| 84 | 70 | 54 | 84 |
| 56 | 74 | 60 | 86 |

What conclusions can you reach from these records?

42. Rates of wear of tires made with two kinds of rubber, A and B, are to be compared by mounting two of each kind on a single automobile. It is known that rates of wear may differ appreciably between front and rear wheels and that no difference is to be expected between the left wheels and the right wheels.

(a) How should the tires be allocated to the four wheels?

(b) What kind of experimental arrangement is this?

(c) After the rates of wear have been measured, how would one form an estimate of the experimental error variance?

43. If, in exercise 42, there is no assurance that left wheels and right wheels produce the same rates of wear, is it possible to make a proper comparison of tires using only one automobile? Can you suggest a better test using two automobiles and four of each kind of tire? What assumption would you have to make to ensure the validity of your comparison of the tires?

44. The effect of the rate of cooling on the hardness of steel can be investigated by selecting several samples from a batch of steel and cooling them at different rates. The effect of composition (e.g. percent carbon), on the other hand, requires whole batches made up for each composition.

An experiment to study both effects is carried out by making four batches, two of composition $C_1$ and two of composition $C_2$. From each batch, three samples are selected, one to be cooled at each of three rates by plunging them into water at temperatures $100^\circ$, $150^\circ$, $200^\circ$. After the specimens have cooled completely, the hardness of each is measured twice.

| | $C_1$ | $C_1$ | $C_2$ | $C_2$ |
|---|---|---|---|---|
| $200^\circ$ | 11.1 | 10.8 | 15.2 | 15.1 |
| | 10.9 | 10.5 | 15.1 | 14.4 |
| $150^\circ$ | 13.1 | 12.0 | 17.7 | 16.1 |
| | 12.9 | 11.1 | 17.6 | 17.3 |
| $100^\circ$ | 15.2 | 14.7 | 20.9 | 18.5 |
| | 15.2 | 14.5 | 20.8 | 19.4 |

Use analysis of variance calculations to sort out the various sources of variation in this experiment and offer whatever conclusions you perceive.