PRINCIPLES OF MEDICAL STATISTICS

VII.—FURTHER PROBLEMS OF SAMPLING: DIFFERENCES

In the previous section the calculation of the standard error of a proportion was based upon a knowledge of the proportion to be expected from some past experience—e.g., if past experience shows that on the average 20 per cent. of patients die, how great a discrepancy from that 20 per cent. may be expected to occur by chance in samples of a given size. In practical statistical work the occasions upon which such past experience is available as a safe and sufficient guide are relatively rare. As a substitute for past experience the experimenter takes a control group and uses it as the standard of comparison against the experimental group. For instance, as the result of the collection of data in this way, we may have the following figures:—

50 patients with pneumonia treated by the ordinary orthodox methods had a fatality-rate of 20 per cent.;

50 patients with pneumonia treated by the ordinary orthodox methods *plus* special method X had a fatality-rate of 10 per cent.

Is this difference more than is likely to arise merely by chance? It is clear that *both* these percentages will, by the play of chance, vary from sample to sample; if method X were quite useless we should, if the results of a large number of trials were available, sometimes observe lower fatality-rates in groups of patients given that treatment, sometimes lower fatality-rates in the control groups, and sometimes no differences at all. In the long run—i.e., with very large samples—we should observe no material difference between the fatality-rates of the two sets of patients; if method X is useless (but innocuous) the difference we expect to observe is, clearly, 0. Our problem is to determine how much variability will occur round that value of 0 in samples of given sizes, how large a difference between the two groups is likely to occur by chance. In other words we need *the standard error of a difference between proportions.*

DIFFERENCES BETWEEN PROPORTIONS

To return to the example previously adopted, let us suppose we take samples from a universe in which is recorded the number of colds suffered by each individual, and for each sample we calculate the proportion of persons who have no colds at all. In the universe itself the proportion of persons having no colds is, we will suppose, 10 per cent. If we take two random samples from the universe, each containing 5 persons, we shall not necessarily observe a proportion of 10 per cent. with no colds in each of these samples; in one sample we may obtain a percentage of, say, 20, and in the other a value of, say, 60, giving a difference of 40 per cent. Is that a difference that is likely to occur by chance in taking two samples of 5 individuals from the same universe? As a test 100 *pairs* of samples of this size were taken from this universe. After each pair had been taken the proportion of the 5 persons in each who had had no colds was calculated, and the difference between these two proportions noted—e.g., in the first sample 1 of the 5 individuals had no colds, in the second sample 2 of the 5 individuals had no colds; the percentages were, therefore, 20 and 40 and the difference between the percentages was 20. This procedure gave the following distribution of *differences between the pairs.*

—	Proportion of persons having no colds in two samples of 5 persons.	
	Differences between percentages in the two samples.	No. of pairs of samples with given difference.
Smaller % in sample A than in sample B.	−60	1
	−40	2
	−20	26
	0	45
Larger % in sample A than in sample B.	+20	22
	+40	4
	+60	—
—	—	100

In nearly half the pairs of samples (45 instances) there was no difference between the percentages having no colds; but there was a considerable scatter round that difference of 0 that we expected to see. For instance in 6 instances the percentage difference was 40 and in 1 it was as much as 60. The mean of the 100 percentage differences is very nearly 0, being −0·6, but very frequently a difference of 20 per cent. was observed between one pair of samples. The scatter of the differences round that mean may be measured, as usual, by the standard deviation; it is 18·0. The distribution of the differences round the mean is, it will be observed, fairly symmetrical, and in a distribution of that shape we know that values that differ from the mean by as much as plus or minus twice the standard deviation are of relatively frequent occurrence. We may therefore say that in samples of this size, in which a percentage of 10 is expected in each, and therefore a difference of 0 between two samples, we may in fact easily observe by chance not that difference of 0 but one as large as ±2 (18)=36 per cent. between the proportions in two samples; differences larger than that will be relatively rare.

Increasing the size of each sample to 20 showed a different distribution of differences between the pairs. The largest difference between 2 samples is now only 25 per cent. (larger differences would be observed if the numbers of samples taken were increased, but these large differences are so infrequent that they are unlikely to occur with only 100 pairs).

—	Proportion of persons having no colds in two samples of 20 persons.	
	Differences between percentages in the two samples.	No. of pairs of samples with given difference.
Smaller % in sample A than in sample B.	−25	—
	−20	6
	−15	6
	−10	15
	− 5	14
	0	27
Larger % in sample A than in sample B.	+ 5	14
	+10	12
	+15	3
	+20	2
	+25	1
—	—	100

The mean of the difference between the 100 pairs is again nearly 0—namely, −1·3 per cent.—but the scatter round that mean as measured by the standard deviation now becomes 9·5; multiplying the size of the sample by 4 has reduced the variability of the differences by half. In samples of size 20, in which a percentage of 10 is expected in each, and therefore a difference of 0 between two samples, we conclude that differences of ±2 (9·5)=19 per cent. between two samples may in fact easily occur by chance, while greater differences will be relatively rare.

Finally taking pairs of samples of size 50 gave the following distribution of differences in percentages having no colds.

—	Proportion of persons having no colds in two samples of 50 persons.	
	Differences between percentages in the two samples.	No. of pairs of samples with given difference.
Smaller % in sample A than in sample B.	−14	1
	−12	—
	−10	4
	− 8	6
	− 6	7
	− 4	2
	− 2	19
	0	6
Larger % in sample A than in sample B.	+ 2	17
	+ 4	17
	+ 6	10
	+ 8	2
	+10	3
	+12	5
	+14	1
—	—	100

In the 100 pairs no difference between the proportions now exceeds 14 per cent. The mean of the 100 differences is +0·9 and the scatter round that mean as measured by the standard deviation is 5·7. In samples of size 50, in which a percentage of 10 is expected in each, and therefore a difference of 0 between two samples, differences of ±2 (5·7)=11·4 per cent. will in fact be relatively frequent and greater differences relatively rare. The standard deviation, or standard error, of the differences decreases, it will be seen, with increasing size of sample.

STANDARD ERROR OF THE DIFFERENCE

The standard error of each proportion is, as shown in the previous section, $\sqrt{\frac{p \times q}{n}}$, where p is the percentage in the universe in one category—e.g., having no colds—and q is the percentage in the other category—e.g., having one or more colds—and n is the number of individuals in the sample. *The standard error of the difference between the two proportions* is, it may be shown, $\sqrt{\frac{p \times q}{n_1} + \frac{p \times q}{n_2}}$, where n_1 and n_2 are the numbers in the two samples. For instance in samples containing 5 individuals drawn from a universe in which p = 10 per cent. and q, therefore, = 90 per cent., the standard error of the difference between the proportion in sample A and the proportion in sample B is $\sqrt{\frac{10 \times 90}{5} + \frac{10 \times 90}{5}}$ = 19·0 per cent. In other words in a single pair of samples each containing 5 persons, drawn from the same universe, we may instead of obtaining the expected percentage difference of 0 quite easily get a difference of ± 2 (19) = 38 per cent. This theoretical value, it will be observed, agrees closely with the value that was obtained practically from the test; the differences found with the 100 pairs of samples of size 5 had a standard deviation of 18. Similarly, the standard deviation of the differences between samples of size 20 was found in the test to be 9·5. The theoretical value is $\sqrt{\frac{10 \times 90}{20} + \frac{10 \times 90}{20}} = 9{\cdot}5$. Finally, the standard deviation of the differences between samples of size 50 was 5·7 and the theoretical value is $\sqrt{\frac{10 \times 90}{50} + \frac{10 \times 90}{50}} = 6{\cdot}0$.

Clearly if we knew the proportion of individuals having no colds in the universe that we were sampling we could calculate the size of differences between two samples that might reasonably be expected to occur merely by chance in taking samples of a given size. If, for example, from the universe used above we took two samples of 50 persons and we treated one sample with, say, vitamin A and found in that sample the proportion of persons having no colds over a specified period of time was 4 per cent., while in the sample not so treated it was over the same period of time 14 per cent., the standard error tells us that that difference is one which might easily arise by chance. The percentage difference between the two samples is 14 − 4 = 10, and this difference, we have seen, has a standard error of 6 per cent. In other words, in taking two samples of 50 individuals from the *same* universe we might easily obtain proportions in the two samples that differed from one another by as much as twice 6 per cent.; a difference of 10 per cent. is not, therefore, a very unlikely event to occur merely by chance with samples of this size. The lower percentage cannot safely be ascribed to the effect of vitamin A for the same difference might quite often occur even if vitamin A were ineffective.

THE STANDARD ERROR IN PRACTICE

In actual practice we do not of course know the value of p in the universe; in calculating the standard error we have to substitute for it a value calculated from the samples. In making this substitution two slightly different lines of reasoning are possible.

(i) We have observed two samples of 50 persons each and the proportion of persons with no colds is 4 per cent. in one sample and 14 per cent. in the other. Is it reasonable to suppose that these two samples have been drawn from one and the same universe in which the percentage of persons with no colds is "x," and that the differences of 4 and 14 from "x" are merely due to chance? Let us adopt the hypothesis that they *are* both samples of this one universe. Then the best estimate that we can make of "x" is given by the *whole* of the observations we have—i.e., the percentage of persons having no colds in our total 100 observations—which is 9. Our question now becomes this: "If we take two samples each of 50 individuals from a universe in which the proportion of persons having no colds is 9 per cent., are we likely to observe a difference of 10 per cent. in the proportions observed in the two samples instead of a difference of 0?" If the answer to this question is *yes*, then we must recognise that though the difference observed *may* be a real one it is quite likely that it is only a chance difference which would disappear if we repeated the experiment. On the other hand, if the answer is *no* we can conclude that our hypothesis that these two samples are likely to be drawn from the same universe is probably not true—i.e., these samples differ from one another by more than is likely to be due to chance and we are entitled to look for some other explanation of the difference between them.

In the example above the difference observed is 14 − 4 = 10 per cent. On the hypothesis outlined above the standard error of this difference is $\sqrt{\frac{9 \times 91}{50} + \frac{9 \times 91}{50}}$ = 5·7. The observed difference between the samples is less than twice its standard error, and we conclude that its occurrence merely by chance is not a very unlikely event—that relatively often we might observe a difference of this magnitude. If the difference had been 15 per cent. we should have concluded that one of this magnitude was unlikely to have occurred by chance, since this difference is more than twice its standard error, and that therefore some cause (perhaps treatment with vitamin A if we are satisfied that the samples were equal in all other relevant respects) had led to the samples differing.

(ii) The alternative approach is to make the hypothesis that the samples are in fact drawn from two different universes; we then test whether the results given by the samples are compatible with the hypothesis that the difference that ought to have been reached in sampling those two universes is zero—i.e., that the two universes are in fact identical. In this case we use as p and q the values found in each sample and the standard error of the difference of 10 per cent. is equal to $\sqrt{\frac{4 \times 96}{50} + \frac{14 \times 86}{50}} = 5{\cdot}6$. As the difference is not twice its standard error we conclude that there is no good reason for supposing that the two universes sampled are different. (Some workers prefer to retain the hypothesis that the samples are drawn from the same universe. We do not know the proportion in that universe but can take the different sample values as two separate estimates of it and use them as above for calculating the standard errors.)

It will be noted that the two methods give very nearly the same results and in practice this is usually the case.

DIFFERENCE BETWEEN TWO AVERAGES

The same type of test is applicable to the difference between two averages, or mean values. For example, the mean height of a group of 6194 Englishmen is 67·38 inches and the mean height of a group of 1304 Scotchmen is 68·61 inches. Are Scotchmen on the average taller than Englishmen or is the difference merely due to chance, inherent in sampling? The standard error of the mean is, it has been shown, $\sigma/\sqrt{n}$, where σ is the standard deviation of the universe sampled and n is the number of individuals in the sample. For this standard deviation of the universe we have to substitute the standard deviation of the sample. The standard deviation of the 6194 Englishmen measured was 2·564 inches, and the standard error of their mean is therefore $2{\cdot}564/\sqrt{6194} = 0{\cdot}033$. The standard deviation of the 1304 Scotchmen measured was 2·497 inches, and the standard error of their mean is therefore 0·069. The standard error of the *difference* between the two means—i.e., the amount of variability the differences might show if we took repeated samples of this size—is $\sqrt{\frac{\sigma^2}{n_E} + \frac{\sigma^2}{n_S}}$, where n_E is number of Englishmen and n_S is number of Scotchmen measured. Unless the observed difference is at least twice this value it might easily have arisen by chance—i.e., the difference between the means ought to be 0 and differs from 0 only by chance. As with proportions we have two alternatives. (i) We may substitute for the standard deviation of the universe the standard deviation of *all* our observations, Englishmen and Scotchmen, put together. The value of this standard deviation is 2·595. By its use we are asking ourselves the question: "Is it reasonable to suppose that we could draw two samples from one and the same universe, in which the standard deviation of the individuals is 2·595, and obtain two means differing from one another by as much as the difference between 68·61 and 67·38?" Inserting this value of the standard deviation we have as standard error $\sqrt{\frac{(2{\cdot}595)^2}{6194} + \frac{(2{\cdot}595)^2}{1303}}$, $= 0{\cdot}079$. The difference between the two means is 1·23 inches, and this is 15·6 times the standard error; it is, therefore, very unlikely that we are drawing samples from the same universe. In other words, Scotchmen are on the average taller than Englishmen—presuming the samples to be representative of the nationalities.

(ii) Alternatively we may use the formula $\sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_S^2}{n_S}}$, $\sqrt{\frac{(2{\cdot}564)^2}{6194} + \frac{(2{\cdot}497)^2}{1304}} = 0{\cdot}076$. In this case the difference between the averages is 16·2 times its standard error. In this test we are presuming that we have drawn the two samples from universes which differ in the variability of their individuals (or from one universe, of the variability in which we have two estimates) and we want to know whether they differ in their means. The two methods give closely the same results.

It must be remembered that none of these methods is applicable to very small samples. Finally a moment's consideration ought to be paid to the level of significance adopted. It has been shown that values, whether of an average or a proportion, that differ from their mean by more than twice the standard deviation are relatively rare. As a conventional level twice the standard error is therefore adopted, and differences between values in two samples which are greater than twice the standard error of the difference are said to be "significant." In fact, differences of this size would occur by chance nearly 5 times in 100 tests. If a worker regards this test as too lenient he can raise his level of significance to 2½ or 3 times the standard error; with these levels, differences would occur by chance, roughly, only once in 80 tests and once in 370 tests. The problem is always one of probability and the worker is at liberty to adopt any level he wishes—so long as he makes his test clear. For this reason it is better to say in reporting results that the observed difference is, say, 2·5 times its standard error, rather than that it is "significant." The latter term will be read as implying that the difference exceeded the conventional level of twice the standard error. The worker who wishes to make the test more stringent may be reminded that he may thereby be classing as chance effects differences that are real; he who tends to be too lenient may class as real differences that are due to chance. With borderline cases there is only one satisfactory solution—the accumulation of more data.

Summary

In practical statistical work the problem that most frequently arises is the "significance" of a difference—e.g., between two proportions, two means, or any other values. The difference between two such values in a pair of samples will fluctuate from one pair of samples to another, and though the samples may be drawn from one and the same universe we shall not necessarily observe a difference of 0 between them. The object of the statistical test is to determine the size of the difference that is likely to occur by chance

(*Continued at foot of opposite page*)

SPECIAL ARTICLES

THE FACTORIES BILL

BEFORE this appears Sir John Simon will have moved the second reading of the Bill to consolidate and amend the Factory and Workshop Acts, 1901–29, and the biggest public health measure of our time will be before the House of Commons. We were able to indicate last week the main changes proposed in existing practice and we may well congratulate those who have thought them out with such care. The Bill contains within it so much that is valuable that those who hope to achieve social progress at one stroke should not permit their anxiety to allow them unduly to obstruct the passage of the Bill into law.

Especially good are the General Safety Provisions (Part II) which are obviously the work of men who know their job. But surely means of access to fire escapes (§35) should not only be marked by a printed notice but should also be kept free from obstruction (as required in §33) ?

Among the General Health Provisions (Part I) §3 aims at securing a reasonable temperature in a workplace and provides the control over braziers, gas fires, and stoves, with their risk of carbon monoxide poisoning, that recent experience has shown is so necessary. It is, perhaps, open to question whether a minimum temperature of 60° F. is high enough to guard sedentary workers against the ill effects of body-cooling. §2 gives power to modify the proposed regulations respecting overcrowding in existing workrooms in which " effective mechanical ventilation " is provided. This might be a dangerous exemption. Is it possible to get effective ventilation in small rooms and shops (apart from air-conditioning) without creating a draught ?

We regret—and our regret will be shared both by industry and the workers—that it was not possible to incorporate in the Bill more exact definitions of reasonable and adequate or effective standards to govern environmental conditions, such as lighting and ventilation. The difficulties no doubt are great but the only alternative is to give such wide powers to the inspectorate as to make their responsibility very great.

§11 gives the Secretary of State power to make regulations for medical supervision in any factory in which excessive illness has occurred, or in which potential risk of illness appears to him to be present. Out of these powers we may hope to see the gradual development of a Health Service for Industry. Since no guiding rules are laid down time alone can show what use will or can be made of this section.

Under General Welfare Provisions §§41–43 govern the provision of washing facilities and the accommodation for clothing. It appears that washing facilities are only to be required where a special need for them exists " by reason of the amount of dust or dirt given off in the process or the dirty or offensive nature of the materials used . . ." Suitable accommodation for clothing not worn during working hours can only be enforced in occupations of a " wet, dusty, dirty or offensive nature . . ." in which workers are " accustomed to remove part of their indoor clothing." We may ask whether either of these two provisions really meets the health needs of a modern community.

§45, which deals with welfare regulations, is no doubt purposely vague. Its effect must depend upon the nature of any new regulations which may be made as a result of its provisions. Is there any reason, however, why arrangements for preparing, heating, and taking meals should not be obligatory, especially where night-work or shift-work is carried on for more than 14 days at a time ?

Coming on to the Special Provisions §51 is welcome, bringing under stricter regulation underground rooms used as factories. §54 gives the Secretary of State power to prescribe the maximum weights which may be lifted, carried, or moved. Medical opinion generally considers that no person ought to carry weights heavier than one-third of his body-weight for any prolonged period of time.

§57 prescribes the conditions under which women and young persons may be employed in certain lead processes. It is hard to understand why these regulations should not apply to all persons (including men) who have to work in an atmosphere containing lead dust. Experience suggests that any concentration of lead above 1·5 mg. per 10 cubic metres of air is potentially hazardous. We should like to see some definite standard stated for these operations and should wish also that every person whose work exposes him to a risk of lead poisoning should undergo regular medical examination. A considerable number of painters still suffer from lead poisoning. They should be as much under control as the accumulator makers. The definition of a lead compound as " any soluble compound of lead " appears to us unsatisfactory. The solubility in water of the oxides or sulphates of lead, for example, may be low, but these compounds are sufficiently soluble in body fluids to give rise to poisoning. Any lead compound which is soluble in the body fluids in-vivo should be included within the definition.

In §119 provision is made for the " medical practitioner who is employed by the occupier in connexion with the medical supervision of persons employed in the factory, who acts as the examining surgeon for that factory, for such purposes as the Secretary of State may direct." We question the wisdom of this provision. It means as it stands, that subject to the authorisation of the Secretary of State, the medical officer (part time or whole time) of a firm may have to act, at one and the same time, in a semi-judicial capacity and also as an employee of the firm.

Under Employment of Women and Young Persons §68 is likely to prove controversial. To permit a 9-hour day and a 48-hour week (or a 54-hour week for over half of the year) for women and young persons is an inadequate restriction of hours. When a five-day week is worked, a 10-hour working day is to be permitted, and this may mean a 12-hour period of employment It seems passing strange that no legal limit should be set to the hours which a man can be asked to work.

But when this Bill has been safely passed—as we hope it will during the present session—large sections

(*Continued from previous page*)

in samples of given magnitudes, how far it may deviate by chance from 0. This involves the calculation of the standard error of the difference. In reasonably large samples this standard error of a difference may be taken to be the square root of the sum of the two individual standard errors of the values in the two samples. A. B. H.