APPLYING STATISTICS IN CRIMINALISTICS*

by

James A. Lechner
National Bureau of Standards
Washington, D. C. 20234

Summary. This paper is intended to tell "why-we-do-it-this-way". After preliminary remarks on randomness, errors, and distribution functions, various techniques of statistical analysis are discussed. These include significance tests, confidence intervals, and goodness of fit tests. Finally, several examples will be discussed: 1) Performance requirements for breath alcohol testers; 2) bivariate discrimination for gunshot residue detection; and 3) Matching "profiles", e.g., trace element analyses or the output of a speech frequency analyzer.

Introduction

Why are there applications for statistics in criminalistics? Because there is randomness in the activities which are the subject matter of criminalistics. And what is randomness? "All the variation, from whatever source, which exists among the results of independent experiments which are intended to be identical" (1). Does a given weapon always leave exactly the same amount of residue on the hand of the firer? Even if it did, would the amount lifted and measured be the same? Why not? How can we predict the behavior of a lift technique? If three different chemists analyze a sample, and come up with three different compositions, what do you do? What if one differs very much from the other two? If you evaluate a breath tester on twenty people, can you make predictions on its behavior in general? What if the twenty people are just picked off the street? Or just came from a businessman's luncheon, complete with cocktails? Or just came from a hospital, and presumably include some on medicine of various sorts? How should one choose a sample?

Some questions have answers so clear that there is little need for statistics. For example, blood type matching: at least among the major classifications, there is no question which group a person belongs to. (Which is not to say that there are never any goofs!) But others require care: if we are to decide whether a person belongs in class A or class B, and we have a measurement, and we know that the class A average is 50, the class B average is 80, and our measurement is 60, what do we decide? But now suppose we have the further information that for class A, the measurements vary from 45 to 55; while for class B, they vary from 50 to 100; anyone care to change their decision? Actually, this is artificial; usually we have not a given finite range, but distributions which extend much farther, and overlap even more, with the spread measured by a "standard deviation". And which decision to take depends not only on the means and standard deviations, but also on the "costs" associated with errors of various kinds, and the proportions of classes A and B in the population or on the "prior odds" that the person in question is from class A. This question I will not pursue any further.

Let us look at another example. Scenario: Sidney Statistician is peacefully working on his latest brainstorm, when Ernie Engineer walks in. Sound:

Hi, Sid - you busy?

Why of course, Ernie, but never too busy to talk to you! What's bothering you?

Well you know, Sid, I'm getting measurements on that infernal new analysis technique, and there seems to be a problem. I got two measurements on each of one control and one suspect, and here's how they turned out:

S	S		CC
15	20	25	x

Well, Ernie, that looks pretty good to me. The measurement technique looks right consistent. What's the problem?

Only one thing, Sid: the "control" was stone sober, so how could he have given higher values than the suspect? I need help!

Okay, Ernie, let's have a look at how the measurements were taken.

Well, Sid, I gave the one batch of two samples to Joe, and the other batch to Bill, and told them to go to it. And I suppose Joe used the one analyzer, and Bill the other; they usually do.

Ernie, you goofed! You have no way to estimate any biases in the equipment or in the operators! Do you have more samples left?

Yes, I do, as a matter of fact. And since I learned something the last time I was here, I have already put two samples from each out for analysis, to a different fellow - i.e., this time Joe got the suspect and Bill got the control, instead of the other way around. And here are the results:

CC			SS
15	20	25	

* This work was supported by the National Institute for Law Enforcement and Criminal Justice, through the Law Enforcement Standards Laboratory of the National Bureau of Standards.

These don't jibe! What do I do now? I only have four samples left.

Let's not cry yet, Ernie. You need to balance things a bit better, if you want to learn anything useful. Suppose you use each of the four combinations of man and machine, once, for each person. That makes four samples each. And just to avoid any personal biases, or time-order effects, let's assign them in random order. OK?

(TIME PASSES. SOME TIME LATER, ANOTHER MEETING OF GREAT MINDS.)

Here you are, Sid - the new data. What do you make of it? Plotted like before, it seems like there might be an effect, but not too clearly:

C	CCS	SCS	S
5	15	25	35

What else can you get out of it?

Let's put the results in a two-way layout. We'll write the suspect value above a slash, and the control value below, for each combination of machine and man, like so:

Man	Machine A	Machine B
Joe	35/25	25/15
Bill	25/15	15/5

(Suspect/Control)

Now notice the consistency: in each cell, the suspect value is 10 higher than the control. And for each combination of machine with sample, Joe is 10 higher than Bill, and finally, for each combination of man with sample, machine A reads 10 higher than machine B. Now I can't say which machine is right, or whether they both have a bias, but it is apparent that machine A's readings are consistently higher; likewise for Joe's readings compared to Bill's. And finally, the suspect's readings are consistently 10 higher than the control's. Good enough?

(CLOSE CURTAIN)

Now anyone knows enough to balance observations like that, doesn't he? (Or does he?) And every experienced scientist knows how accurate his technique is, right? Wrong! Consider Fig. 1. It shows fifteen measurements of the astronomical unit (the mean distance from earth to sun), made

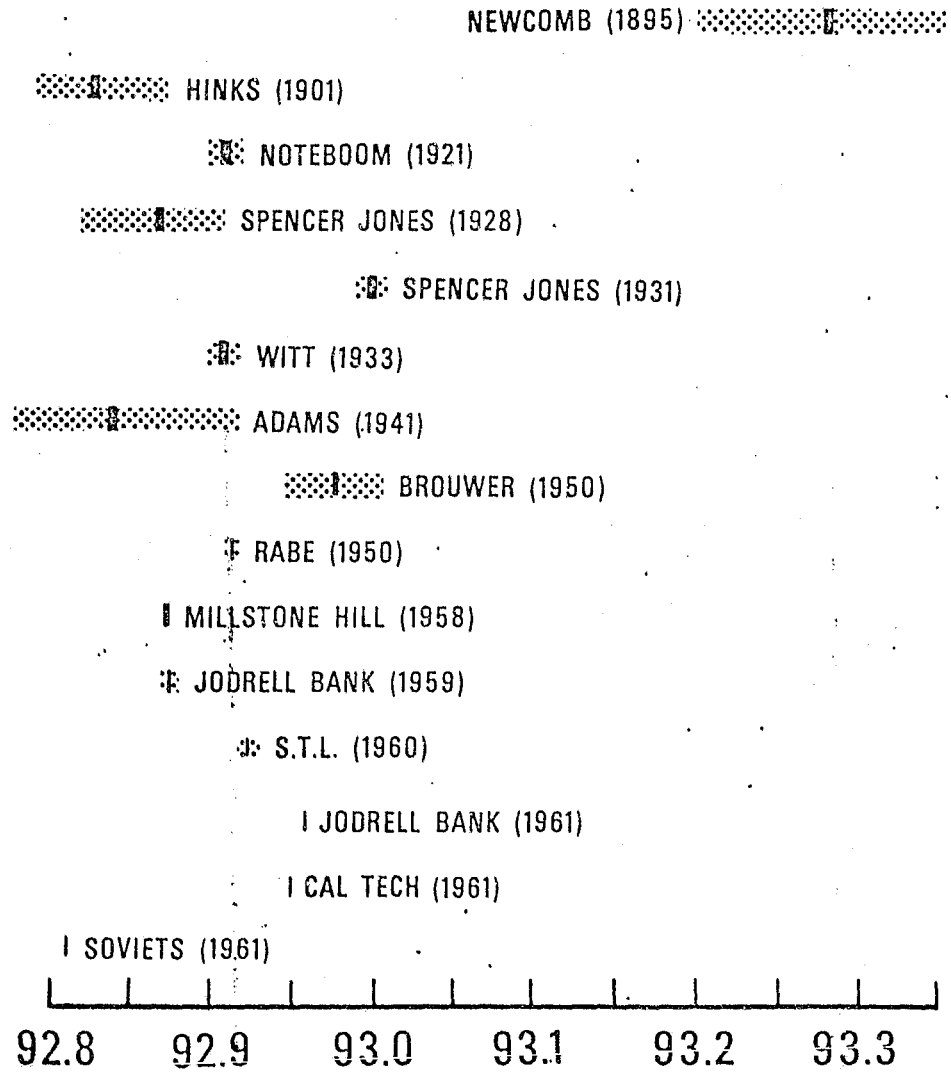


Fig. 1. Estimates of the astronomical unit (mean earth-sun distance), with uncertainties. Each value is outside the range of uncertainty of the preceding value.

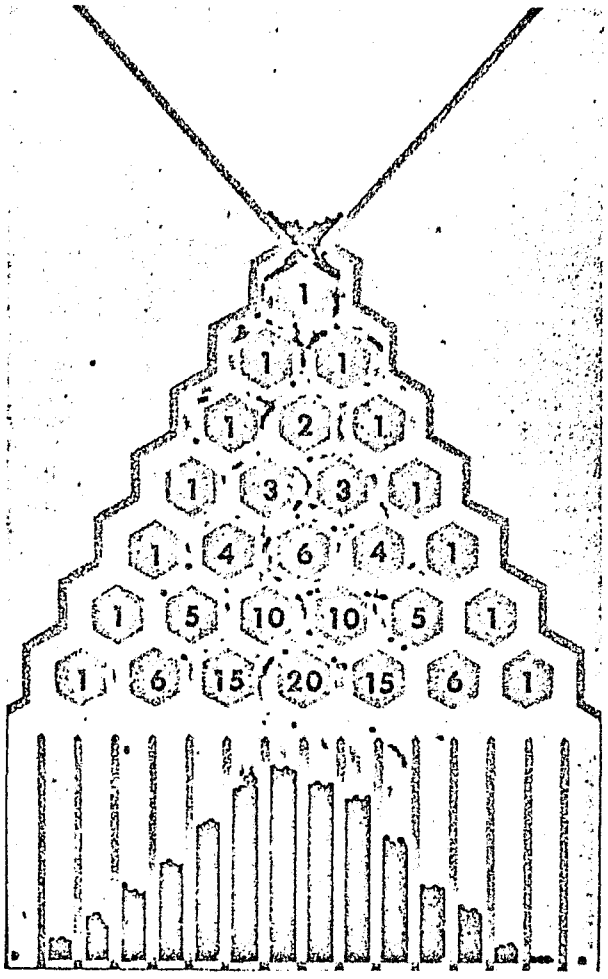


Fig. 2. The Quincunx, or Galton probability board. The balls will approximate a Binomial distribution, if the board is properly constructed. The numbers constitute a "Pascal triangle"; each row gives the relative frequencies for a binomial distribution with p = 1/2 and appropriate value of n.

over the period 1895 to 1961, by highly respected scientists (2). Also shown is the uncertainty assigned to each measurement by the scientist(s) who produced it. The striking thing about this set of measurements is that not one of the measurements fell within the uncertainty assigned to the previous measurement! There are probably several morals to be gotten from this, but one I would like to state is: Don't take for granted that any component of error is negligible - measure it!

Not much fancy statistics went into the example above. But the experimenter was fortunate: the experimental error was so small relative to the effects, and the "design" was so bad, that the problem came to light. Rest assured, there are many times where it does not - it just results in false positives, or false negatives, or no conclusions at all, and the loser may be a manufacturer, a suspect, the public at large, a researcher, or - you name it. How much better to anticipate the sources of error, and measure them. They don't cease to exist by being ignored - they just get to foul things up in ways that never can be righted! So call in the statistician early, and be open with him!

Errors

There are several well-recognized sources of error. One classification is into the following:

Variability of experimental material

Smaller batches tend to be more homogeneous. (And this applies whether the batching is by process lot, by time, by distance, operator, producer, or anything else.) If you stick to one batch, you might get pretty results - but do they apply to anything other than that batch? Remember, the purpose of experimentation is generally to make inferences about other material!

Uncontrolled conditions

One does not always know what external conditions influence the results. (Temperature? Day of the week? Whether it is payday? Etc.) Even if one does know, one may choose not to control - or may not be able to control. But at least, one is aware that there is a source of variability there, and is not likely to make assertions about the results that could be invalidated by that source of error.

Measurement error

This includes human error. Again, don't be optimistic. Take an honest look at the variability.

There is one technique of such general usefulness that it ought to be mentioned, even in so general an essay as this. That is the technique of comparison experiments. Suppose that one has a measurement technique that is subject to many kinds of errors, which cannot be well controlled. Suppose however that the effects of these errors are the same for all specimens. Then the difference between two specimens would be pretty much the same under any set of conditions, as long as the two specimens were measured under identical conditions. Finally, suppose that we have standards - i.e., items whose characteristics are well known. Then we can measure the test specimen and the standard, keeping all experimental conditions as constant as possible, to get a good estimate of the difference. Knowing the true value for the standard, we can then adjust the measurements to get a good value for the test specimen. (And of course, there are refinements to evaluate how constant the conditions were kept, involving repeat measurements.) This approach is not a cure-all, but it is a big help in many situations.

Distribution Functions

A distribution function tells us how we think something behaves. (It's a model.) It tells us how the results of many trials (independent trials, that is) would stack up. For example, Fig. 2 shows what is known as a Quincunx, or Galton probability board. Each number on each row is formed by adding the two numbers directly above it; these numbers constitute the "Pascal triangle", which

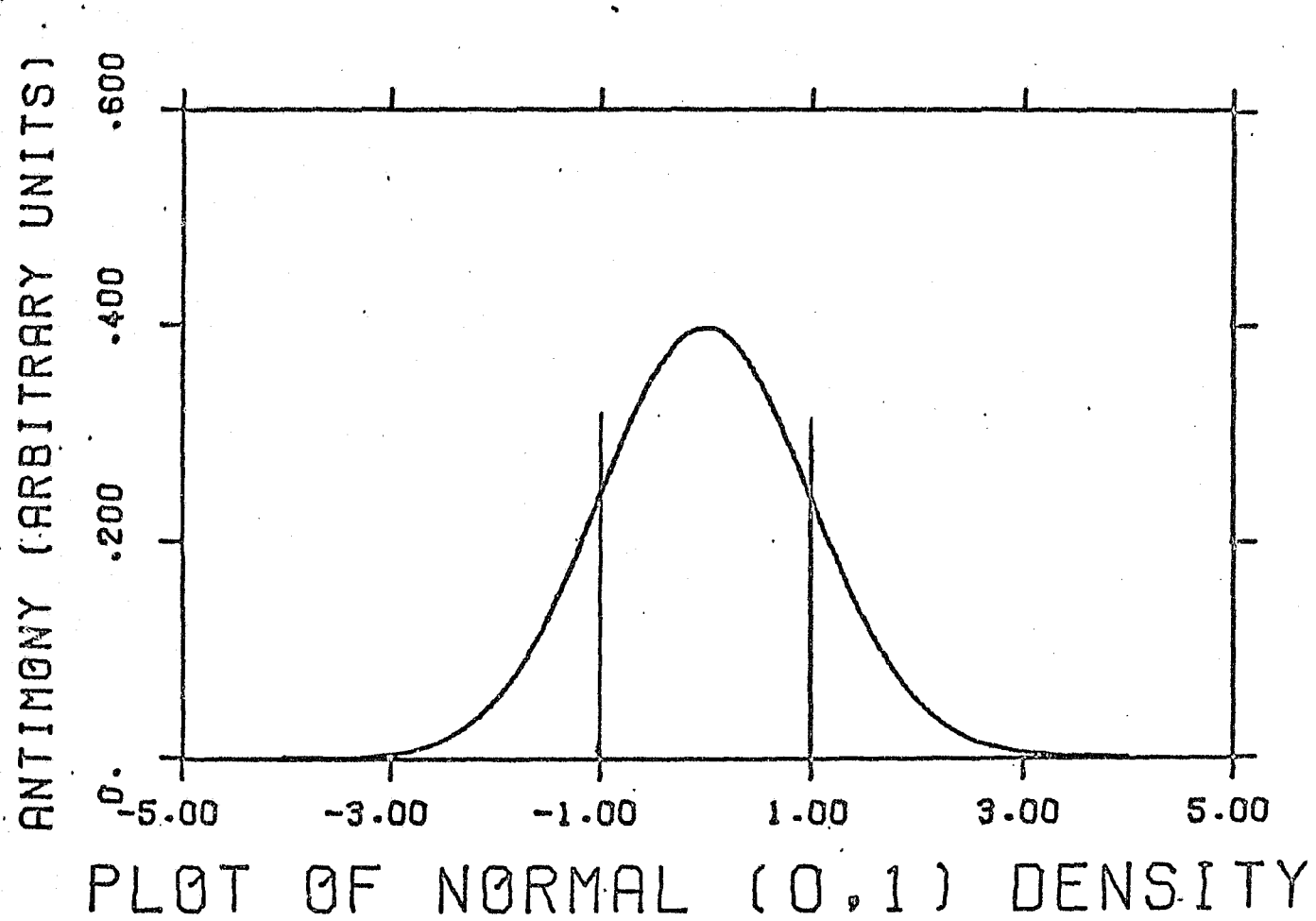


Fig. 3. The Normal distribution with zero mean and unit variance (the "standard Normal"). The vertical lines enclose 68% of the probability.

gives the relative probabilities in the binomial distribution with $p = \frac{1}{2}$, for each value of n. For example, if a fair coin is flipped five times, the number of heads has a binomial distribution with $n = 5$ and p (the probability of a head on one toss) $= \frac{1}{2}$. Thus, looking at the sixth line (the first corresponds to $n = 0$), we see that the relative frequencies of 0,1,...,5 heads is 1,5,...,1. Since the sum of this row is 32, the probabilities are 1/32, 5/32, ..., 1/32. A little reflection will show easily why this is true, if you consider what happens to a lot of marbles which have a 50-50 chance of falling to the right or to the left of each peg. Now within sampling variability, the shape of the pile of marbles accumulated at the bottom should correspond to a binomial distribution. But it will also look familiar to many of you as a "Normal", or Gaussian, distribution. (Everybody believes in the Normal distribution - the experimentalists, because they think it is a theoretically proven fact, and the theorists, because they think it is an objectively-observed reality!) There's some truth in both views: it has been shown that, under the right conditions, the sum of a large number of variables (e.g., the contributions from many different sources of error) tends to behave like a Normally-distributed variable, and on the other hand, so many physical measurements seem to have a Normal distribution. BUT NOT ALL! The "right conditions", alluded to above, are satisfied when the number of components is large, their contributions are independent, and the contribution of any one is small relative to the total. This does not include, for example, the case when there are occasional bloopers, or "outliers". Nor does it include the case when many components are correlated - i.e., tend to be large or small together.

Fig. 3 shows a Normal distribution, with mean equal to zero and variance equal to unity. (The so-called "standard" Normal.) Now suppose this is the appropriate error distribution for some measurement, and we want to know the true value to within 1/10 of a unit. One measurement is not going to give us an answer we can trust. Everyone knows what to do - take a number of measurements and average them. Why does this help? Because when measurements are Normally (and independently) distributed, then the mean of a set of n is also Normally distributed, but with variance smaller by a factor of n. So if we take 100 measurements, and average them, the mean will quite likely be within 1/10 of the true value - in fact, since its variance is 0.01, its standard deviation is 0.1, and the tables of the Normal distribution tell us that a variable will fall within one standard deviation (in this case, 0.1 units) of its mean, about 68% of the time. If that is not enough, take more

measurements: with 400 measurements, the s.d. is 0.05, and the probability of being within two standard deviations (again, 0.1) of the mean is 95%. But remember, this is true for independent observations only! If the apparatus is not taken down each time, the observations are not independent: some of the errors which affect each measurement are being held constant throughout the experiment, and thus will appear in the mean just as strongly as they do in a single measurement! (They become randomly-chosen "bias" errors instead of random errors.)

Now let's consider tests of significance, and confidence intervals. We will use a simple example, because it's the concepts we are after. Suppose we have a coin, and we wish to test whether or not it is "fair". We wish to decide between the two hypotheses H_0: "the coin is fair" and H_1: "the coin is biased". A very important point to realize here is that these two hypotheses are different in kind: if H_0 is true, we know exactly the behavior pattern of the coin, and can predict (for example) what will be the distribution of the number of heads in a large number of tosses; while if H_1 is true, the behavior pattern is not completely specified, because we don't know how biased the coin is. A corresponding situation holds in agricultural experiments, where significance testing got its start: H_0 is often a hypothesis of no difference in effect between different "treatments", for example. In fact, people often try to find such a simple H_0, even if it entails some Procrustean efforts. Why? So we have something to hang our hats on, so to speak: we need to be able to analyze the test procedure. We are generally interested in seeing the null hypothesis rejected, but not (of course) when it is true, and primary emphasis is often on limiting the chance of rejecting a true H_0.

Back to the coin. If we perform n tosses, and count the number of heads, then we should have approximately n/2 heads, if the coin is fair. If the number of heads is too far removed from this value, we will reject the hypothesis of fairness. How far is too far? That depends: what happens if we err? Note there are two kinds of error possible: rejecting a true H_0, and failing to reject a false H_0. (Note I did not say "accepting

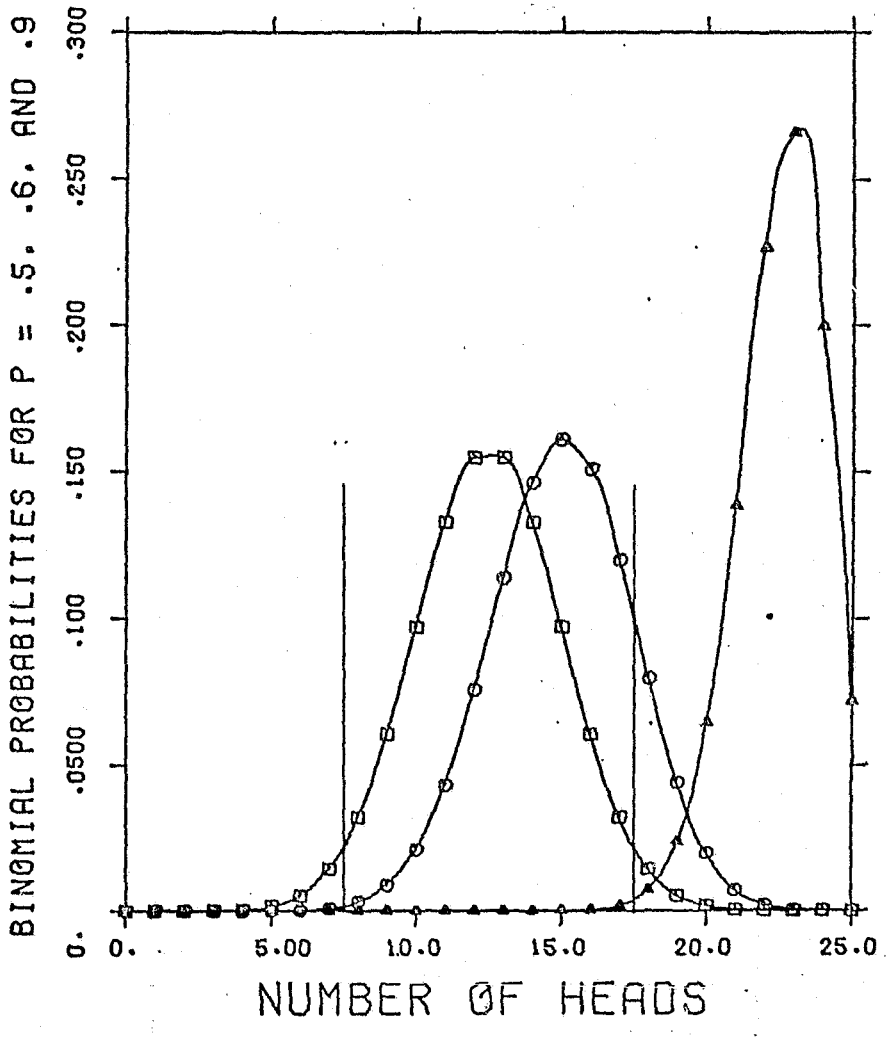


Fig. 4. The Binomial distribution, for n = 25, p = 0.5, 0.6, 0.9. The probabilities to the right of the vertical line at 17 1/2 are 0.022, 0.153, and 0.998 respectively.

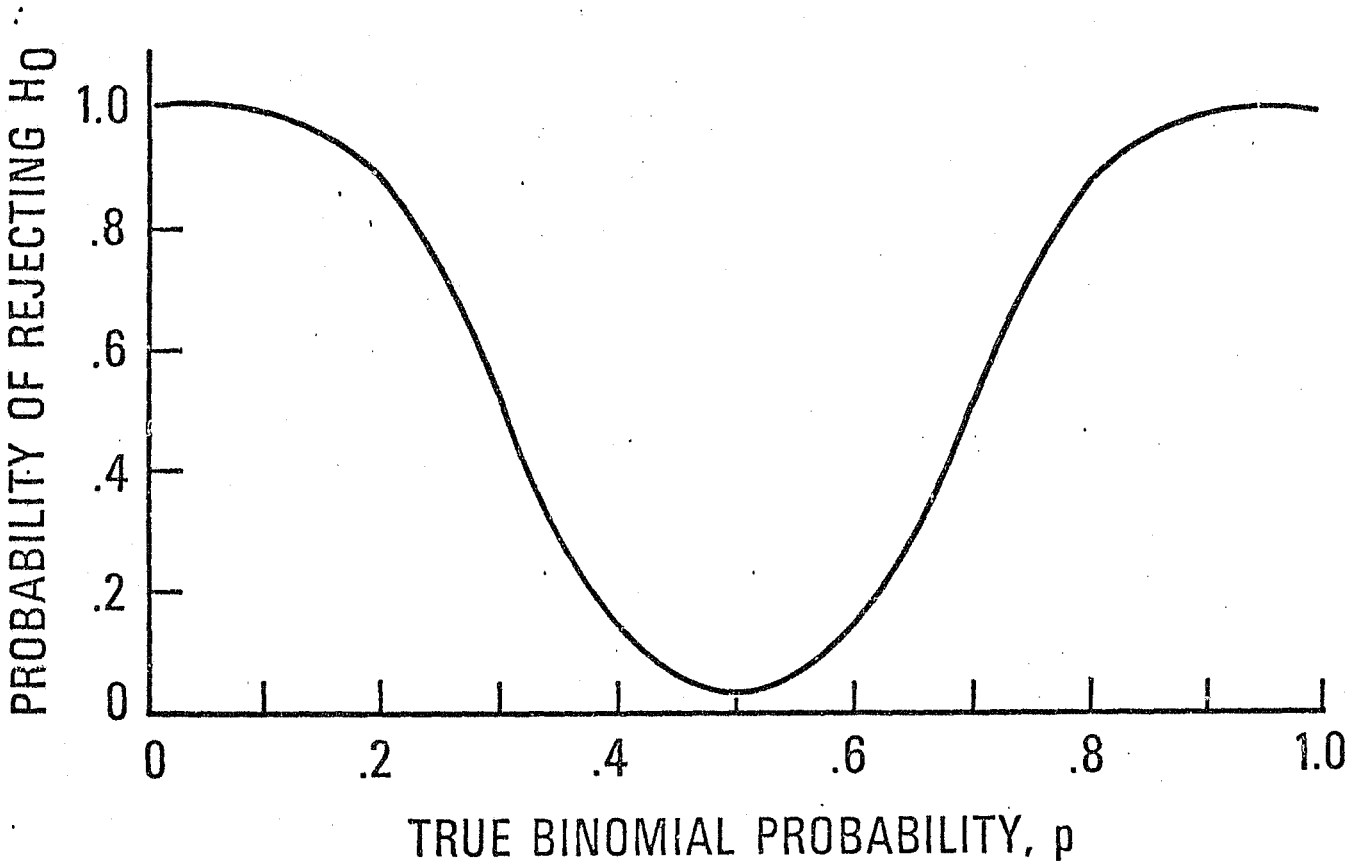


Fig. 5. The "power" (probability of rejecting H_0) for a two-sided 5% test of H_o: p = 1/2) for the Binomial distribution with n = 25.

a false H_0". More about this point later.) Since under H_0 we know all about the behavior of the coin, it is comparatively easy to set a value for the first type of error (which incidentally is called the Type I error): given the truth of H_0, the distribution of the number of heads is binomial with known parameters, and so we can calculate the probability of getting a number of heads which is beyond whatever criterion we set. For example, if we decide to reject whenever the number of heads differs from n/2 by more than $n^{1/2}$, we have (approximately) a 5% chance of rejecting a true H_0. (Here I have used the Normal approximation to the binomial distribution, so it is valid for n greater than about 30; a correction for continuity would allow its use for somewhat smaller values of n.) Thus we can settle the question of safety: we have limited the chance of incriminating an innocent coin. What about effectiveness? In other words, what chance do we have of incriminating a biased coin? Look at Fig. 4: it shows the binomial distribution for $n = 25$, and $p = \frac{1}{2}$, with a cutoff (criterion) value of $\pm n^{1/2} = \pm 5$ around the expected number of $12\frac{1}{2}$. Also shown are the distributions for p = .6 and p = .9. Obviously the chance of rejecting a coin with a 0.1 bias is small: only 15%. If it were desired to have a large chance of rejecting such a coin, while preserving the 95% safety level, we would have to perform more flips - in fact, we would need 270 flips to attain a 90% chance of rejection. (The derivation of this result is left as an exercise for the reader.)

Statisticians often talk about a "power curve". This is merely a graph of the probability of rejection (the power of the test) as a function of the parameter of interest (in this case, the true probability of heads). The power curve for the test above, with n = 25, is shown in Fig. 5. The power curve is bound to be low at the parameter value corresponding to the null hypothesis, since we want the "size", or Type I error, to be small. It will generally rise from that point (at least in the direction of the alternatives of interest). We would like it to be steep, however, so that even slight departures from H_0 would have a good chance of being caught. As is often the case, steeper curves cost more. Therefore it behooves one to think carefully about what alternatives are likely, and how important it is to catch them, before experiments are done!

Incidentally, it is now obvious why we don't ever say we have "accepted" H_0. If the results of the experiment are highly surprising under H_0, we can reject H_0, and we have the assurance that in so doing we will reject a true H_0 only once in so many tries - whatever corresponds to the error level we are using. But if the results are not that surprising, all we can say is that we do not reject H_0; it may be because we did not do a big enough experiment, or because H_0 is not as far

wrong as we thought it might be, instead of being proof that H_0 is true.

Now a word or two about confidence intervals. We just said that the number of heads occurring in n tosses of a fair coin is likely to be within $n^{1/2}$ of the expected value n/2. As a matter of fact, this can be extended to any coin that is not too biased: the number of heads occurring in n tosses of a not-too-biased coin has a 95% chance of being within $2n^{1/2}$ of its expected value, which is np, where p is the probability of a head for that coin. In symbols, $\Pr(np - n^{1/2} < x < np + n^{1/2}) = 0.95$. But by a little trivial algebra, we note that the double inequality above is equivalent to the following: $x/n - n^{-1/2} < p < x/n + n^{-1/2}$. Can we then say that there is a probability of 0.95 that the true p falls between the two limits $x/n \pm n^{-1/2}$? Yes, if we are careful about how we interpret such a statement. Note that p is not usually a random variable, but rather it is a constant, even if it is unknown. Therefore it either is or is not between the two calculated limits, and the probability is accordingly either 1 or 0. The real meaning is that there is a probability of .95 that two limits, so calculated, will in fact bracket the true value. The chance of confusion is sufficiently great that different terminology has been adopted: we say we have 95% "confidence" that the true value lies between the quoted limits. The practical meaning is that, in a long series of such statements, only one in twenty times will the quoted interval fail to enclose the true value.

Finally, a few words about goodness of fit tests. Fig. 6 shows a (hypothesized) true distribution function, and a sample cumulative distribution function. They differ, of course. Is there enough evidence to say that the sample probably does not come from the distribution shown?

There exist various tests for this problem. Some are based on the maximum difference between the two curves; some on the integrated area between the two curves; and some on the difference between actual and predicted frequencies in a set of intervals covering the range of possible values. One of the latter is the so-called Chi-squared test. It is indeed widely used: anytime there is (or can be produced) a classification scheme, the Chi-squared test is likely to be used. If there are n cells, or intervals, and the theoretical distribution predicts values E_i (which stands for Expected) for cell i, and one actually observes O_i, then it is a simple matter to calculate $(E_i - O_i)^2/E_i$, sum over all i, and compare with the Chi-squared tables, using n-1 for the degrees-of-freedom parameter. It often works fairly well. Why? Because in many applications, the distribution of the calculated statistic is reasonably close to a Chi-squared distribution. But there are a lot of approximations going on here. And

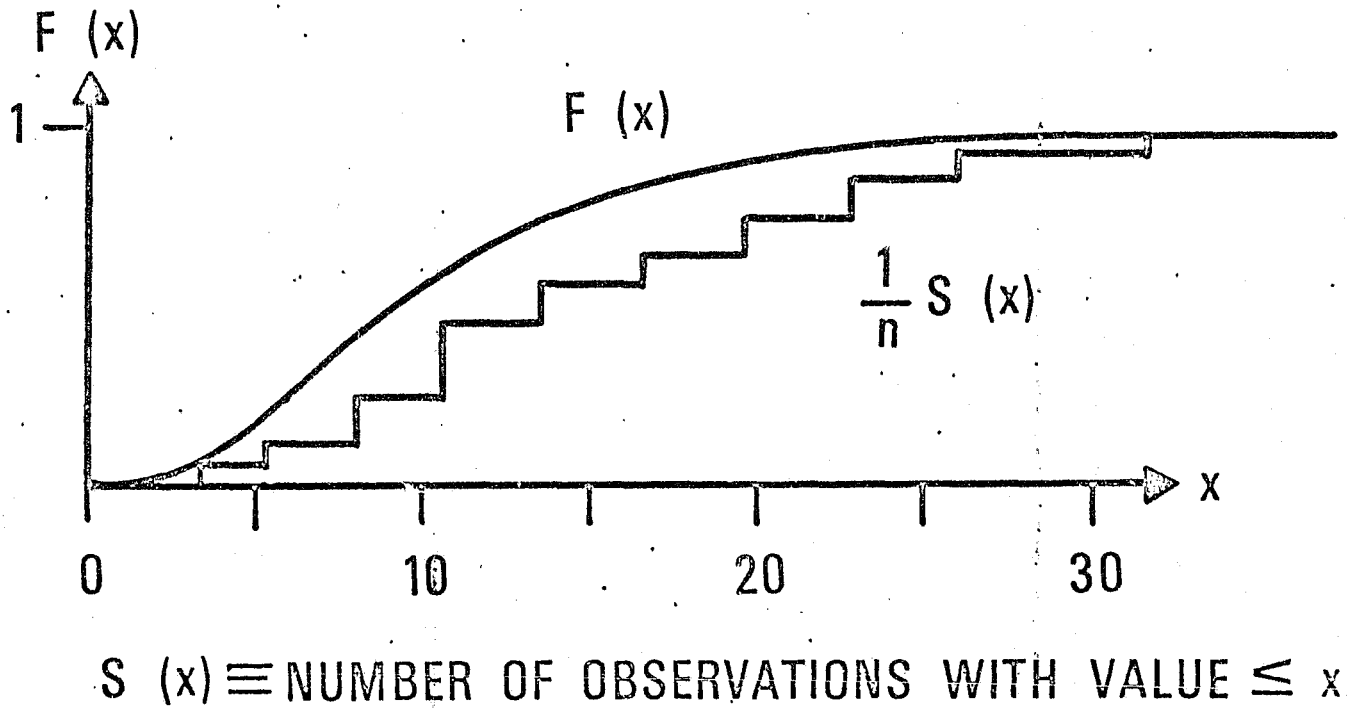


Fig. 6. Continuous cumulative distribution function (cdf) F(x), and sample cdf $n^{-1}S(x)$. (S(x) is the number of sample values $\leq x$; n is the sample size.)

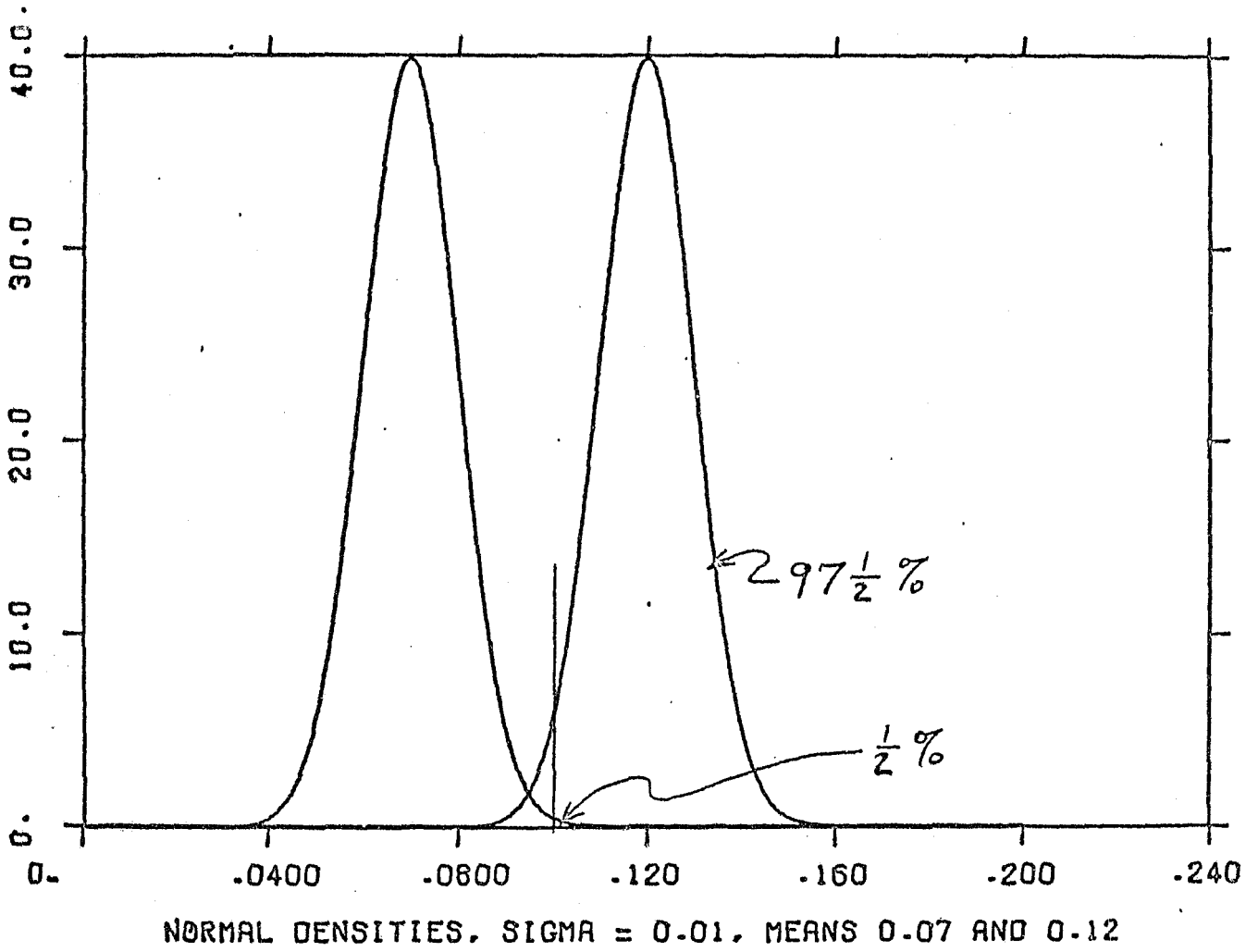


Fig. 7. Normal distributions, for mean values of 0.07 and 0.12, with standard deviation equal to 0.01.

since anything that can go wrong will go wrong, people get into trouble all the time by ignoring them. Furthermore, it has happened that people have used the Chi-squared when there are no frequencies in sight. What happens then? One case in point will be mentioned in the third example of the following section, where more detailed discussion of Chi-squared appears.

Examples

Standards for a breath analyzer: acceptance criteria

Let's start with some background information. In most states, the legal limit for blood alcohol concentration is 0.10% w/v. NBS has developed a standard for breath alcohol testing machines, which (among other things) requires that the standard deviation of readings, as estimated by a standard test procedure, be at most 0.004% w/v. There has been some criticism of this level, to the effect that it ought to be lowered to 0.003% w/v. The effect of such a change was to be evaluated.

From experiments at NBS, it seems that the standard deviation of machines currently being produced is about 0.0028% w/v. It is not a good idea to set a requirement so tight that satisfactory machines cannot pass, and going a little farther, it is also not a good idea to set the requirement so that the machines can pass (with reasonably high assurance) only after a very expensive test. To be more specific, if the true s.d. were in fact .0028%, and we felt the manufacturer was entitled to have at least 95% of such machines pass the test, then we should set the limit and the number of tests performed on each machine so that there is at most a 5% chance of having an estimate, based on that number of tests, higher than the limit. Basing the test on the sample standard deviation, one needs 9 or 10 observations with a limit of 0.004, but about 280 with a limit of 0.003! Of course, one also wants to be sure that machines which do get accepted are satisfactory; i.e., that machines which are not satisfactory have a very low chance of acceptance. It turns out that, with ten observations, and with a s.d. of 0.01, the chance of acceptance is less than 1/2 of 1%. And since 0.01 is still much smaller than the legal limit of 0.10, it still provides a perfectly adequate margin of safety for the innocent and a good chance of detection for the drinker: with a true value of 0.07, there is only a 1/2% chance of getting a result above 0.10, while

with a true value of 0.12, there is a 98% chance of getting a result above 0.10. (See Fig. 7.)

Bivariate discrimination for gunshot residue detection

Problem: Given values for the amounts of barium and antimony found on the (firing) hand of a suspect, decide whether the values are incriminatingly high: i.e., whether they provide evidence that the suspect has recently discharged a weapon. Preliminary thoughts: Do we want to pick a threshold level for each element, say T_a for antimony and T_b for barium, and then treat as a "positive" result any case for which either value exceeds its threshold, or only those for which both values exceed their respective thresholds? Should they depend on time since firing is suspected to have occurred? On type of weapon? On the particular weapon, if known? On sex, or occupation, or number of shots, or seriousness of the crime, or type and brand of ammunition? On lifting technique? Or lab, or lifter, or wind conditions?

Null hypothesis: Suspect is innocent. (Denoted by H_0.) We want to bound the probability of a "Type I error", namely the rejection of H_0 when it is in fact true. (At what level? People often pick 95%, but why? Would you be content with a 95% probability that any given hydrogen weapon will not go off accidentally, if there are 100 of them scattered around your city? I hope not: there's only a 0.6% chance that there will not be a disaster! It makes sense, oftentimes, to iterate this decision: let's assume a tentatively chosen value, say 95%, and go on. This means that out of every 20 persons who have not recently fired a weapon, but who are tested, one (on the average) will be wrongfully accused (on the basis of this procedure). What then happens with persons who have fired a weapon? If it turns out, upon investigation, that in order to keep the Type I error small enough, the threshold has to be so high that we obtain a positive result for only one of every 10 who have fired, what do we do? We can work to make the procedure more sensitive; we can lower the threshold; we can be satisfied that occasionally we will produce a positive; but we can no longer delude ourselves into thinking we have a good procedure as it stands.

Now let's look at the situation graphically. Fig. 8 shows a bivariate distribution of barium and antimony levels. It is to be interpreted as a contour map: there is a "lump" of probability, of unit volume, which is spread out over the (x,y) plane, forming a "hill". If we were to integrate

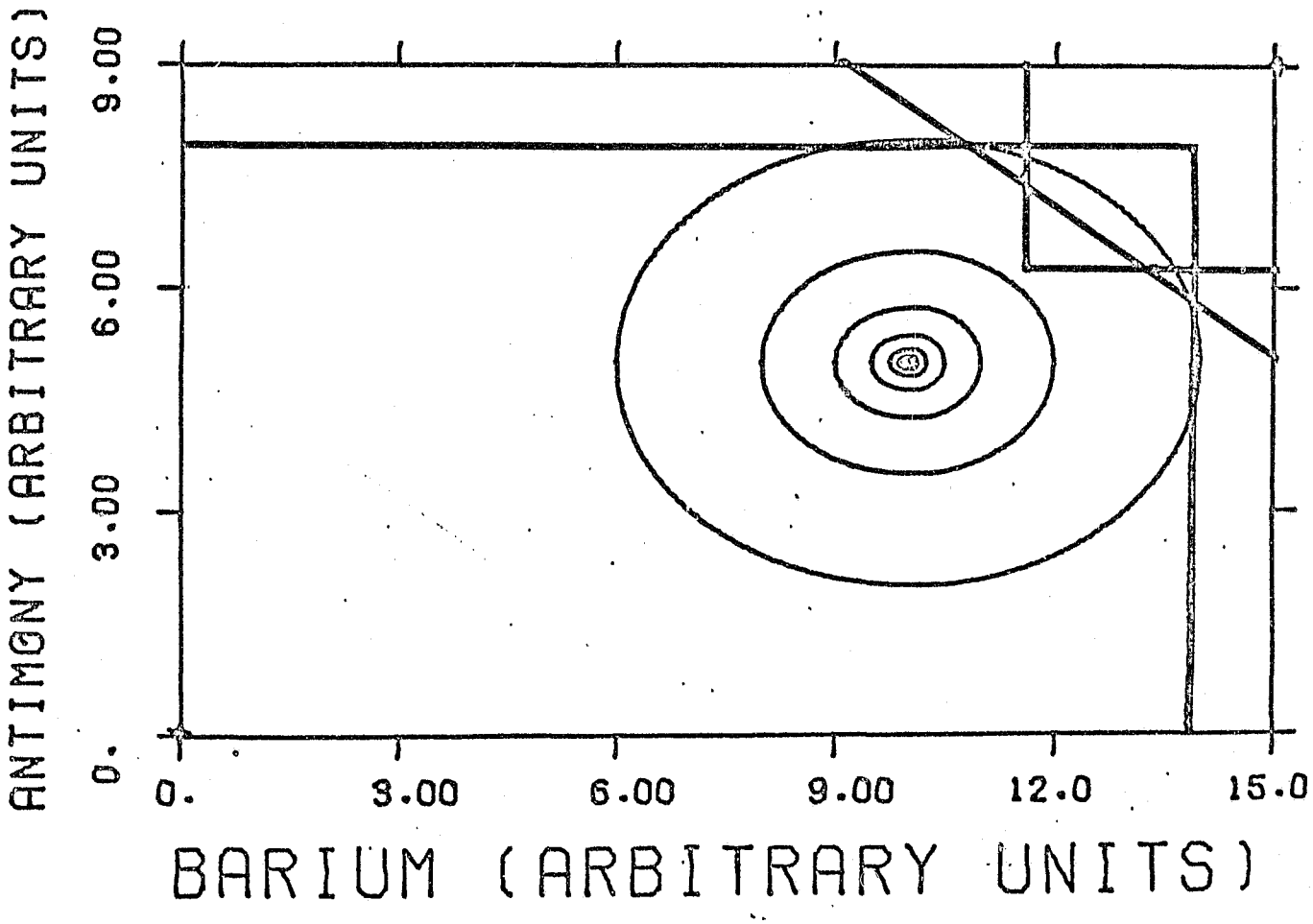
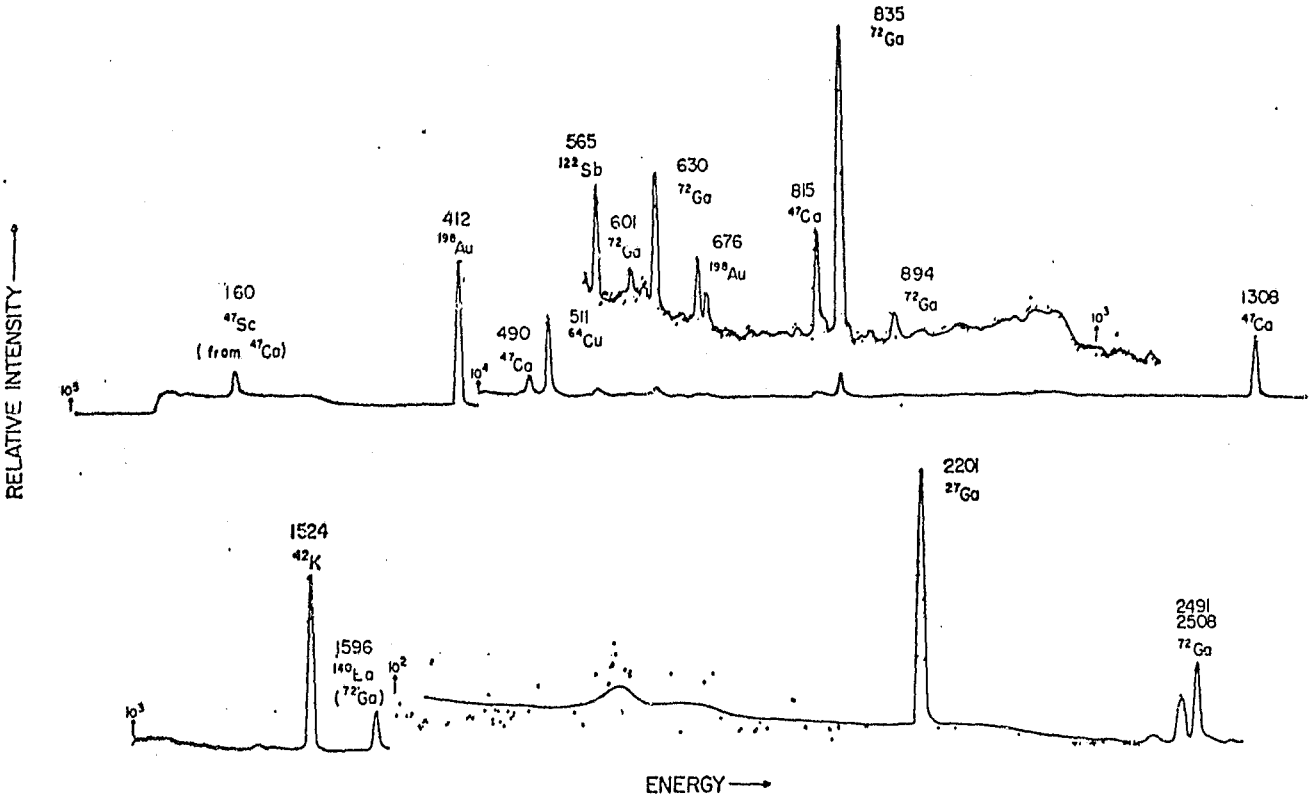


Fig. 8. Bivariate Normal distribution with possible cutoff criteria. (See text.)



Gamma-ray spectrum of HAP effluent of irradiated glass sample; 100 min counting time; 48 hr after irradiation.

Fig. 9. Typical Neutron Activation Analysis output curve.

the elemental volume in this hill, with respect to the variable y, we'd have a probability density for x; and vice versa. As drawn, the figure does not indicate any appreciable correlation between x and y; if there were a correlation (i.e., a situation where knowing the value of x changes the distribution of y), it would show up as a tilt in the axes of the elliptical contours.

A horizontal line on the figure represents a threshold level for y; a vertical line represents a threshold for x. If we are to keep the Type I error to 5%, then we need to pick these thresholds accordingly. If we are only going to test x, and not y, then the vertical line needs to be drawn so that 5% of the volume of the hill is to the right of it. If only y, then the horizontal line needs to be drawn with 5% above it. If x and y are independent, then we can use lines with $2\frac{1}{2}$ % above and to the right, respectively, declaring a positive whenever either line is exceeded. (i.e., whenever the measurement is outside the lower left quadrant.) Or we could use lines marking off 22%, and demand that the actual values reach the upper right quadrant before declaring a positive. Or perhaps a slanted line is best. The point is, we don't know which is best until we know what values are likely to occur for actual firers! All the above are equivalent, as far as true negatives are concerned, and we want to pick a rejection region which includes as many as possible of the true positives. And there are complications: the values for actual firers change drastically with ordinary human activity over time, and depend on the weapon, and are extremely variable even for the same weapon; and the values for non-firers depend on occupation. It is reasonably certain that with certain ground rules, the technique is indeed valuable; but its efficient use demands study beyond any yet reported in the literature. It is not a simple problem!

Matching profiles: multivariate discrimination

When a trace element analysis is performed, one typically obtains a plot of "power" against frequency, as shown in Fig. 9. This then must be converted to a table of percentages for the trace elements present, as in Fig. 10. Certain techniques of fingerprint analysis involve measuring and/or counting various characteristics of the prints, and also result in vector observations somewhat like Fig. 10. If one has such sets of values for a number of samples, or one sample and a set of "standards", how does one tell which ones are alike? If we could visualize 15-dimensional space, it might help. Various ad-hoc techniques have been applied to this kind of problem. Each has adherents, but nothing very definitive is known.

One technique in use for comparing two samples, or one sample with a known population, involves calculating a quantity which it is hoped behaves like a Chi-squared variable, and comparing it with the appropriate statistical tables. This approach is fraught with danger, however. To see why, let us look at the Chi-squared quantity. It is usually used when there are a number of

categories, and each of a large number of items falls (independently of the other items) into one of the categories. One has an expected number for each category, obtained by assuming each is equally likely, or from some other assumption that seems reasonable (and is to be tested). One takes the squared difference between observed and expected numbers for each category, divides by expected, and sums over all categories. This number is (under certain assumptions) reasonably like a Chi-squared variable with parameter n-1, when there are n categories. To see why this is reasonable, turn to the definition of a Chi-squared variable: the sum of squares of k independent variables, each one having a Normal distribution with mean zero and variance unity, is Chi-squared with parameter k. How does this match up with the categories situation? Well, the number of items which actually get classified into a certain category is a Binomial variable (assuming that each item independently gets classified into one of "many" categories, according to fixed probabilities); but if the probability associated with a given category is "small", and the total number of items classified is "large", then the number which end up in the given classification is (approximately) Normally distributed with variance equal to the mean. Thus if we subtract the expected (mean) value, and divide by the square room of the expected value, we have (approximately) a Normal with mean zero and variance unity; upon squaring, we have the usual Chi-squared formula. But how many categories, how small a probability, and how large a number of items? And what should we do if we are comparing two samples, and have no reason to consider either one the "expected" one? There are several rules of thumb for the first question. One says that every expected frequency should be at least 2, or that at least 80% of the expected frequencies are at least 5, with the remainder being at least 1; with the exception that when the parameter (the number of "degrees of freedom") is 1, all expected frequencies must be at least 5. The other question has a more definite answer: one calculates expected numbers based on row and column totals, in a layout where (say) rows represent the different categories, and columns represent the different samples. The proportion of the grand total represented by the i-th row is an estimate of the (assumed common) probability associated with the i-th category, and so when we multiply it by the j-th column total (the number of items in the j-th sample), we get the expected number in the (i,j)-th cell. Again we take observed minus expected, square, divide by expected, and add over all cells; only now the parameter is (r-1)(c-1) where r and c are the number of rows and columns respectively. Again the restrictions on minimum values for the expected numbers must be observed. Remember, what is being tested here is that the distribution among categories is the same for each sample. If rejected, then one has evidence that the samples do indeed come from different populations. Now one more thing needs to be said. Suppose that one actually has percentages of trace elements, instead of counts in various categories. Can one force this situation into the Procrustean mold of a Chi-square analysis? Not without doing violence to both the situation and the technique! Percentages are not counts, even if they are obtained from a radiation counter! (It may be that the scaling, combined with the accuracy of the measurement process, combined with Heaven knows what else, counterbalances nicely to make the result look like a Chi-squared variable. But I certainly would not like to stake my life on it, without checking it out. Nor would I like to stake my freedom on it, so why should I stake somebody else's freedom on it?) This problem is being investigated at the moment, with the hope that some (at least partially satisfactory, interim) techniques will be forthcoming.

References

1. Hooke, Robert, "Introduction to Scientific Inference", 1963, Holden-Day, San Francisco.

2. Youden, W. J., "Experimentation and Measurement", 1962, Scholastic Book Services, New York.

Table I. Results of Base Glass Analysis, ppm[a]

Mn	0.479 ± 0.014	(±2.7%)
Au	0.217 ± 0.013	(±6.0%)
Cu	1.035 ± 0.030	(±2.9%)
Ga	0.330 ± 0.010	(±3.0%)
La	0.0291 ± 0.0015	(±5.1%)
Sb	0.0643 ± 0.0041	(±6.4%)

[a] Limits quoted are $ts/\sqrt{n}$ for the 95% C.L.: $n = 18$.

Fig. 10. Trace element composition, determined from a curve similar to Fig. 9.

END