The Bell System Technical Journal

January, 1924

Relays in the Bell System

By S. P. SHACKLETON and H. W. PURCELL

NOTE: Before they can converse people must either be brought together or virtually be brought into one another's presence by the telephone. Any telephone system must establish talking connections between its subscribers, and these connections must be built up, supervised and disconnected when desired. This work is accomplished by the use of relays of various kinds, and the speed and accuracy of service is largely dependent upon them. There are completed daily in the Bell System about 42,000,000 telephone calls. These involve the successful and accurate operation of over one and one-half billion contact connections daily.

Many kinds of relays are employed in the Bell System, varying from the simple electromagnetic drop to the sequence switch, the thermionic vacuum tube and the panel selector. Today a circuit connection between two subscribers served by manual exchanges in a large multi-office district involves about 21 relays. When these subscribers are served by machine switching offices, the number of relays in a local connection may be as great as 146. It not infrequently happens that in setting up a toll connection more than 300 relays are employed.

In the present paper the relay developments leading up to, and making possible the present communication system, are outlined with particular reference to electromagnetic relays. A few typical circuit applications are given with a discussion of the requirements imposed upon relays which influence their design. Several types of relays are illustrated and their distinctive features are described.

The subjects of relay design, manufacture and maintenance and also telegraph relays will be dealt with in future papers.

INTRODUCTION

IN the vast systems of networks which comprise the Bell System one of the most important and varied devices necessary for giving service is the relay. From its use in small numbers in telegraph circuits and as a "drop" in the early magneto switchboard it has come to be numbered by millions and varies in type from the simple electro-magnet which operates a single contact to the vacuum tube and the complete structure which effects an entire series of switching operations.

When a small number of stations is involved in a communicating system complete flexibility of connection may be obtained by means of simple relays controlling a small number of contacts. As the number of stations increases the number of switching operations becomes so great that the use of simple relays which control small numbers of contacts is not economical. The use of power driven selectors and sequence switches and electro-magnetically operated switches for completing a series of switching operations has therefore become necessary.

In present day systems the relay is as essential to a telephone conversation as the transmitter or receiver. Some idea of our dependence on the device may be had from a consideration of the numbers of simple relays involved in a typical connection. A circuit established between two subscribers served by manual exchanges in a multi-office district will involve 21 relays. When these subscribers are served by machine switching offices the number of relays

Fig. 1—Relays in a local manual office

involved in a local connection may be 146. When toll connections are involved even greater dependence is placed on relays to render service. A New York-San Francisco connection requires over 200 relays and very frequently connections are established which require more than 300 relays. The majority of these relays are normally available for doing their bit to provide telephone service to any one of a large number of subscribers. As a matter of fact, approximately 90 per cent of the millions of relays in the Bell System today are available for and may be called upon to serve any subscriber or user of the telephone.

A typical manual office serving 10,000 lines would have from 40,000 to 65,000 relays and their total combined pull if applied at one point would be sufficient to lift ten tons. In the larger machine switching offices there may be as many as 140,000 relays which require in some instances power plants capable of handling peak loads of 4,000 amperes at 48 volts.

Referring to Fig. 1, the space required for mounting some of the relays in an office will be seen. This is a picture taken in one of the New York offices which has over 60,000 relays and the racks shown contain about 22,000 of these. The covers have been removed from a number of the relays in the foreground. Instead of grouping the relays compactly as in a manual office it is the practice in machine switching offices to mount them in close association with the related apparatus units. This is illustrated by the photograph of sender circuit relays shown in Fig. 2.

Invention of the Electromagnet

Prior to 1820, the electro-magnetic structure, now known as a relay, was an impossibility because the scientific facts on which it is based had not been discovered. In the winter of that year, Oersted of Copenhagen established that a mechanical effect could be produced on a magnetized needle by a current of electricity. Oersted discovered that a magnetic needle would be deflected from its normal position when held parallel to a wire conveying an electrical current and that the deflection would be to the right or left, according to the direction of current flow. This discovery aroused such interest among scientists and philosophers that the best minds in Europe were engaged in speculation and experiment, so that further discoveries of great importance followed rapidly. Arago in Paris and Davy in London, working independently, soon observed that, if an electric current passed through a wire of copper or any other material, the wire had the power of inducing permanent magnetism in steel needles.

Oersted's discovery suggested to Schweiger that the mechanical effect on the magnetic needle would be increased if the current were made to pass several times around the needle. He made a coil, elliptical in shape, of insulated wire and suspended the magnetic

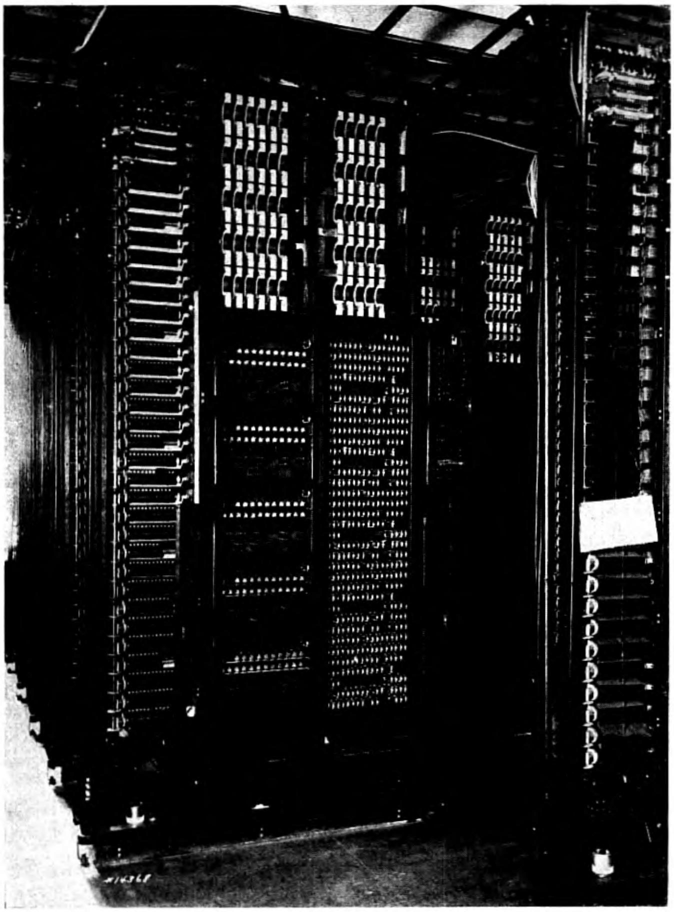


Fig. 2—Sender relay cabinet in machine switching office

needle within it. When current passed through the coil, the result was as he anticipated, and the device became known as Schweiger's multiplier.

Ampere, the brilliant French scientist, in seeking an explanation for Oersted's discovery, evolved an ingenious theory of the relation between electricity and magnetism. According to this theory, all magnetic phenomena result from the attraction or replulsion of electric currents supposed to exist in iron at right angles to the length of the bar, and all the phenomena of magnetism and electro-magnetism are thus referred to one principle—the action of electrical currents on one another. Among other things, he proposed a plan for the application of electro-magnetism to a system for transmitting intelligence. This system was to operate by the deflection of a number of needles at the receiving station by currents transmitted through long wires. By completing a circuit the needle was to be deflected and was to return to normal under the influence of the attraction of the earth when the circuit was opened. This proposed system of Ampere's was never reduced to practice, however.

All these discoveries and results were prior to 1823, and they resulted in the development of needle telegraph systems, which were at one time employed extensively in Europe. These systems utilized a coil of wire around a magnetic needle pivoted in the center and with a pointer attached to the needle, which was suspended over a dial. Deflections to the right or left signified letters and were produced by sending pulsations of one polarity, or alternations of both, as was required.

In 1824, Sturgeon, an Englishman, discovered that, if a current of electricity flows in a coil of wire surrounding a bar of annealed iron, the latter becomes a magnet, and when the current ceases, the iron loses its magnetism. Sturgeon bent an iron rod into the form of a horse-shoe and wound a coil of copper wire around it loosely, with wide intervals between the turns to prevent them from touching each other. Through this coil, he transmitted a current. The iron under the influence of this coil became magnetic and thus, the first electro-magnetic magnet, now known simply as the electro-magnet, was produced. This discovery of Sturgeon's is of great interest to the telephone and telegraph engineer, because it was the direct step which made the invention of the electro-magnetic relay possible.

In 1828, Henry, in America, after repeating the experiments of Oersted, Ampere, Sturgeon, and others and investigating the laws of the development of magnetism in soft iron by means of electrical currents, designed the most powerful electro-magnet that had ever been made. This he accomplished by associating Schweiger's multiplier with Oersted's magnet. For this purpose he wound 35 feet of silk insulated wire around a bent iron bar so as to cover its whole length with several thicknesses of wire.

First Telegraph Relays

In 1824, Morse utilized the electro-magnetic phenomenon, revealed by Sturgeon, and produced a telegraph system which was destined to be the basis of all modern systems of communication. The attenuation of the current from the sending to the receiving end of the circuit had limited the satisfactory transmission of signals. Morse overcame this difficulty by constructing an electro-magnet which would repeat or "relay" the transmitted signals to another circuit having an independent source of energy. The first electro-magnet or relay designed by Morse was a cumbersome structure weighing about 300 pounds, but it exerted a tremendous influence on the art of communication as it served as a stimulus for the development of the complex systems of the present day. When this relay was redesigned its weight was reduced to 70 lbs., but as the laws of electro-magnetism became more generally understood and new materials became available, such great changes occurred that the present telegraph relay weighs about 3 pounds, and one of the modern telephone relays of latest design weighs but 3.6 ounces.

The General Problem

The needs of the present day telephone and telegraph system have produced a multitude of devices but none of them is of greater importance than the relay, as it affords the means whereby the engineer may put ideas into practice. When the limitations of available relays prevent the satisfactory solution of a problem, requirements for new relays are outlined and their development is undertaken if a survey indicates that the advantages to be obtained warrant the expense.

This does not mean that compromises are not made in the matter of using standard designs, for in some instances, it would not be economical to design a new type. Frequently, a relay is required to meet certain conditions in the plant for which the demand will be comparatively limited, and it is obviously uneconomical to spend time and money developing a new type provided a standard structure can be adapted to meet the requirements with sufficient precision.

Just as the art of defense in warfare has matched the art of offense, the art of relay design has kept pace with the demands of the circuit engineer. Relays are now required to operate on direct, and pulsating current, and also on alternating current throughout the entire range of frequencies which are used in communication. There are fast relays, slow relays, polarized, high impedance, low impedance and so on. Consequently, a relay designed for one purpose may be wholly unfit for any other use. On this account, as the telephone art has grown, new conditions and new requirements have resulted in the development and manufacture of a large number of relays. At first, this undoubtedly followed previously established precedents, so that new forms were brought into existence which fulfilled immediate needs, but did not receive much consideration from the standpoint of economy, standardization or consistency of design.

At the present time, the Western Electric Company manufactures for the Bell System about 100 types of simple electro-magnetic relays. These types are subdivided into about 3,500 kinds, which differ in minor ways, such as windings and contact arrangements. In 1921, the Western Electric Company produced over 4,800,000 of these relays. These figures serve to indicate the economic importance of the relay in the present day system but do not give any adequate conception of the dependence of the communication network on relays of all types.

From a design standpoint it is possible, as has been pointed out, to attain practically any desired result in an electric circuit, subject of course to certain limitations as to time, and provided no limitations as to economic application are to be met. The methods and means for securing the desired operations involve the use of relays of various types and designs, and may lead to new developments which are obviously not economical. The relay may be called upon to perform a single function, necessitating the opening or closing of a single contact, or it may be required to effect a complicated series of transfers or circuit changes. Its operation may necessitate an accurate synchronizing with other circuit operations involving a time lag in its operation or release, and other requirements as to impedance, power, etc., may be imposed. It frequently happens that the conditions imposed by circuit requirements necessitate a choice between new features of relay design and a complication of the circuit to overcome limitations in existing types of relays.

The economic considerations which govern the final application of circuits in the telephone system are, to a large extent, dependent on the costs and performances of the various types of apparatus, particularly the relays. Frequently, there may be a number of possible methods of accomplishing a given result in an electric circuit and the most economical method is, of course, desired. This does not necessarily indicate the least number of relays or the cheapest but rather the most economical combination, taking into account reliability of circuit operation and its effect on service, cost of equipment and cost of operation and maintenance. The more complicated circuit or the one necessitating additional equipment may be sufficienty more reliable to justify its use.

In considering the application of relays in any telephone circuit, a given problem is usually presented and the various possible methods of accomplishing the desired end are considered. These methods may involve combinations of relays or of relay parts which do not exist and may even involve combinations which are entirely impractical or uneconomical of application. Any simplifications which may be effected are considered and in case the design of any apparatus may effect appreciable savings in the circuit or otherwise appear justifiable, this may be undertaken. Such requirements on relay design are, of course, subordinated to any general design considerations, such as relay structure, etc., which may be governing from the standpoint of the economic production of the relays themselves.

A few considerations which influence the selection of relays and which are very closely associated with the fundamental relay design may be considered from the standpoint of their effect on telephone circuits and their application in the field. It would, of course, be impossible within the scope of this paper to describe all the relay applications in modern telephone practice, but a discussion of the relay requirements for a few typical cases will serve to illustrate the principles involved. While the first relays used in the telephone system were telegraph relays adapted for use in signaling, the vast majority of relays now in use in the Bell System are designed primarily for telephone circuits. The requirements are usually quite different, particularly as regards the energy available for operation, the speed of operation and reliability of contacts and in most cases the cost.

Early Telephone Relays

In the first telephone switchboard for commercial service which was installed in 1878, the electro-magnetic devices consisted of a telegraph relay and an annunciator for each subscriber's line, and a call bell common to all lines. Of these three the telegraph relay was the largest and most costly, so the desirability of reducing the number required and of providing a smaller and cheaper apparatus unit was apparent. Changes were soon made in the magneto system that removed the relay from the subscriber's line and associated each relay with a group of lines for supervisory purposes. In the early switchboard, patented in 1879, from which the standard switchboard was developed, a modified telegraph relay appeared as a clearing out relay to control a clearing out drop. This modified telegraph relay is shown in Fig. 3 and is of particular interest as representing the first step in the development of the telephone relay.

In the magneto systems the indicator or drop was of the first importance, so its development was rapid. It was finally arranged in one extensively used system with two coils, which were known as

Fig. 3—Early telegraph relay used as telephone drop

the line coil and the restoring coil. The magneto current from the subscriber's station energized the line coil to drop a shutter which was restored through the agency of the restoring coil when the operator inserted a plug in the associated jack. The early development of the drop undoubtedly influenced the forms of relay structures which were developed a little later. The analogy between the line and restoring coils of the magneto system and the line and cutoff relays of the common battery system is very close. In the latter the current over the subscriber's loop energizes the line relay which lights the line lamp. The insertion of the plug in the subscriber's jack energizes the cutoff relay which opens the circuit through the line relay and thus extinguishes the light. In addition, the line and cutoff relays are assembled on a common mounting plate, forming an apparatus unit, although they are not parts of the same structure as were the corresponding coils of the drop.

The early telephone relays resembled more closely in construction and form the early drop than they did the telegraph relay, although the influence of design work on the telegraph relay appears in the development of later types of telephone relays. The early telephone relay shown in Fig. 4 has little relation with the modified telegraph relay of Fig. 3. It is much smaller and lighter than the first relay and, in addition, there is a distinctive structural change in that the armature is suspended by a reed hinge.

At this time the limiting conditions controlling the operation of either telephone circuits or the apparatus in them had not been established with much precision, so that the requirement for a relay was, roughly, that it should do the work required and any arrangement more sensitive and reliable than a previous arrangement was an improvement. The principle of the reed hinge for an armature support was sound, in that it provided for a good magnetic circuit and an easy means for close air gap adjustment and it is now used extensively with relays of the latest design.

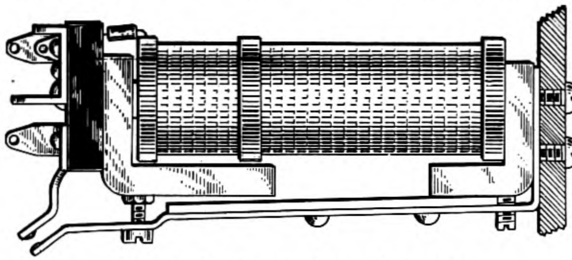


Fig. 4—Early telephone drop relay

Line, Cutoff and Supervisory Relays

When the common battery system was developed, however, it was found that the reed hinge relay was not capable of meeting the additional requirements imposed by the new system. The common battery cord circuit originally suggested is shown in Fig. 5. It is apparent that the relay shown in the diagram must indicate positively to the operator the position of the switch hook in the substation set and must respond to the motion of the switch hook if the subscriber moves the hook up and down to interrupt the circuit. In addition, as this relay is in the transmission circuit it must not introduce objectionable transmission losses. The reed hinge relay could not meet these additional requirements, and accordingly a new relay was designed especially for the common battery system that was the most important single factor in making the new system possible. In order to obtain an armature that would respond quickly to any change in the holding magnetic force all forms of support for the armature were rejected. The relay developed is shown in Fig. 6

and was first used in the common battery board installed in Worcester, Mass., in 1896.

This relay consists of a tubular magnet with an iron disc armature in the form of a truncated cone. This disc is brought to an edge at its periphery and rests in an annular groove in the cap. When the armature operates, it closes against an insulated contact stud

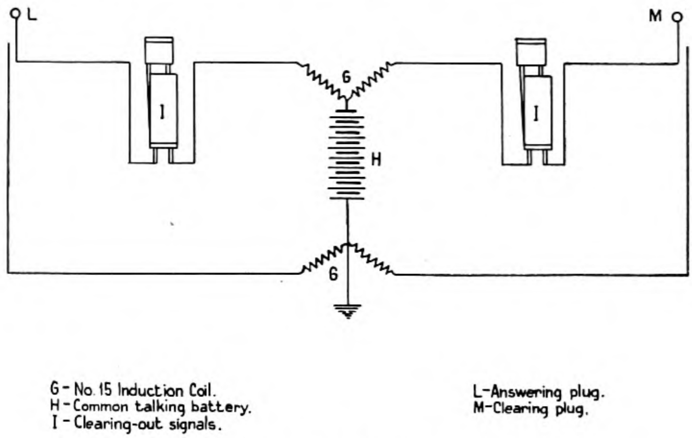


Fig. 5—Early common battery cord circuit

projecting from the core and when released drops away from the core by gravity and rests against a stud projecting from the end of the cap which provides the adjusting means for regulating the armature air gap. As the contacts of this relay were enclosed in the case, they were protected from dust and this arrangement proved so desir-

Fig. 6—Early line relay

able that it has been an accepted feature of nearly all relay designs that have followed. This arrangement had the disadvantage, however, of not providing a means for determining the value of the armature air gap or the contact separation. This condition was improved in the next design which is shown in Fig. 7 by making the cap longer and associating the armature with the magnet structure instead of the cap. The knife edge armature hinge and the gravity control of the free armature were the fundamental principles retained in the new design.

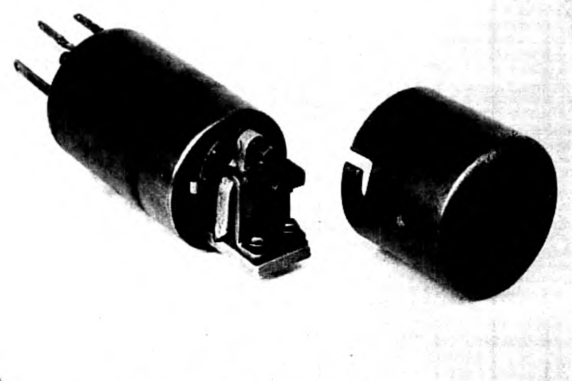


Fig. 7—Early supervisory relay

The tubular shield for the return magnetic path was abandoned for a return pole piece which provided a means for mounting both the armature and the air gap adjusting screw. The cover protected the contacts from dust but it was soon found that magnetic interference between adjacent relays was responsible for both faults in

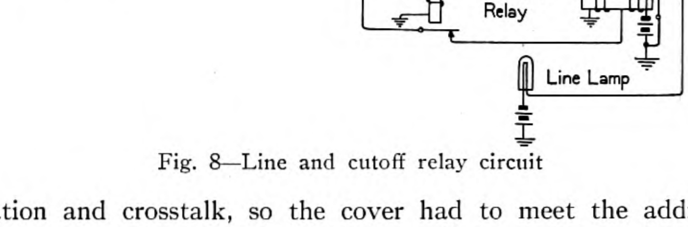


Fig. 8—Line and cutoff relay circuit

operation and crosstalk, so the cover had to meet the additional requirement of being an effective shield. This was eventually accomplished to best advantage by making the cover of copper.

As has been shown two general types of structures were now available for common battery system relays. In one, the armature was suspended by a reed under tension to provide a restoring force. In the other, which was more sensitive but less capable of carrying a heavy spring load, the armature operated in a knife edge hinge and the restoring force was gravity.

Each subscriber's line required a l'ne relay for lighting a lamp when the substation receiver was removed from the hook, a cutoff relay for removing the line relay from the circuit when the operator responded, and a supervisory relay for controlling lamp signals to inform the

Fig. 9—Line and cutoff relays

operator whether the substation switch hook contacts were open or closed. The circuit arrangement for the line and cutoff relays is shown schematically in Fig. 8.

The rapid extension of telephone service necessitated establishing standards of excellence, and definite requirements for apparatus units were gradually formulated. At first, the available relays were adapted as closely as possible to existing conditions, but as requirements became definitely established, relays were designed specifically to meet them and careful consideration was given to manufacturing costs, mounting space, maintenance expense and all other factors of economic importance. By 1910 several million of the line and cutoff relays shown in Fig. 9 and the supervisory relay of Fig. 7 were in service.

The cutoff relay armature was of the reed hinge type, while both the line and supervisory relays were assembled with the more sensitive knife edge armature. The line relay was eventually wound with 12,000 minimum turns to a resistance of 2000 ohms $\pm$ 5 per cent. and after considerable service experience requirements were formulated for a line relay which would be a satisfactory substitute. These requirements were as follows:

(1) Battery potential, 20–28 volts.

(2) Maximum line resistance, including subscriber's station, 1000 ohms.

(3) Resistance across line to represent maximum insulation leakage, 10,000 ohms.

(4) Winding of relay 12,000 turns, 2000 ohms $\pm$ 5 per cent.

(5) Minimum operating ampere turns =

$$\frac{\text{Turns} \times \text{Minimum Voltage}}{\text{Maximum Resistance}} = \frac{12{,}000 \times 20}{1000 + 2100} = 77.4.$$

(6) Maximum releasing ampere turns =

$$\frac{\text{Turns} \times \text{Maximum Voltage}}{\begin{array}{c}\text{Leak Resistance} + \text{Minimum}\\ \text{Relay Resistance}\end{array}} = \frac{12{,}000 \times 28}{10{,}000 + 1900} = 28.2.$$

Reference to the circuit will show that the line relay must release on a low resistance loop in case the subscriber flashes the line lamp to attract the operator's attention. Due to residual magnetic effects, a relay does not release after operation on short loops over which the operating current is high as quickly as it does after operating on long loops, with a lower current in the winding. It is, therefore, necessary to specify the maximum ampere turns the line relay may receive and adjust it to release immediately afterward with the maximum leak across the line.

(7) Maximum ampere turns $= \dfrac{12{,}000 \times 28}{1900} = 176.8.$

In addition:

(a) The relay must close one set of contacts which controls a signal lamp as shown in the circuit.

(b) Contacts must carry the energy for lighting the lamp without undue sparking, sticking or wear.

(c) The relay must operate reliably on 77.4 ampere turns.

(d) The relay must release on 28.2 ampere turns immediately after operating on 176.8 ampere turns.

(e) As there is a constant potential between windings, the coil must be protected from corrosion, so the materials chosen for the construction of the relay must not contain substances which tend to encourage or assist electrolytic action.

It was found that the line relay introduced high maintenance charges because of the knife edge armature hinge and the close adjustment required. The armature being comparatively light in weight, a slight amount of dust or corrosion in the armature slot frequently made the contact resistance in the slot so high that the line lamp would fail to light.

The supervisory relay in the cord circuit also introduces maintenance charges for the same reason. It had to meet the same requirements as the line relay but, in addition, it required a crosstalk proof cover that would also adequately protect the contacts from dust.

The solution of this problem was very difficult because the only means of obtaining relays of increased sensitivity or greater operating range was to make mechanical refinements, which would increase manufacturing costs quite out of proportion to the advantages obtained, to discover new magnetic materials of higher permeability at low flux densities and with a lower remanence characteristic or to develop an entirely new relay structure. To obtain any advantage from the development of a new structure built from the same magnetic materials, it would be necessary to design it in such a manner as would enable the engineer to obtain the proportion of copper and iron required for maximum efficiency, greatest economy, extreme sensitivity, maximum operating load or any other specific requirement which was the controlling factor of a particular design. Analytical studies had shown that smaller relays with less iron could be substituted for those in use but such a change could not be made without increasing manufacturing costs because a reduction in core diameters would increase breakage during manufacture as well as entail a greater cost of handling the smaller structures.

The Flat Type Relay

An analysis of the manufacturing costs had shown labor costs to be greater than material costs in the production of relays so that any changes which would result in large savings would have to be of such a nature as to reduce labor charges. This could be accomplished only by changing production methods which had already been established with reference to greatest economy in manufacture considering the volume of production. The demand for relays, however, was increasing steadily and it was evident that with increased production the prevailing manufacturing methods would not continue to be economical. With other pieces of apparatus manufactured in large quantities, it had been found that production costs could be reduced to a minimum by designing a unit which could be assembled from interchangeable parts stamped out by a punch press and formed in bending fixtures to the desired shapes. To accomplish this for relays, it was first necessary to conceive of an entirely new type of structure composed of parts that could be made easily by the punch press method, and it was then necessary to determine the modifications which must be made in such a structure, in order that the proportions of copper and iron required for any one of a number of design conditions might be obtained.

The design of a punched type relay was first attempted in an effort to find a better relay than the line relay, and with the intention,

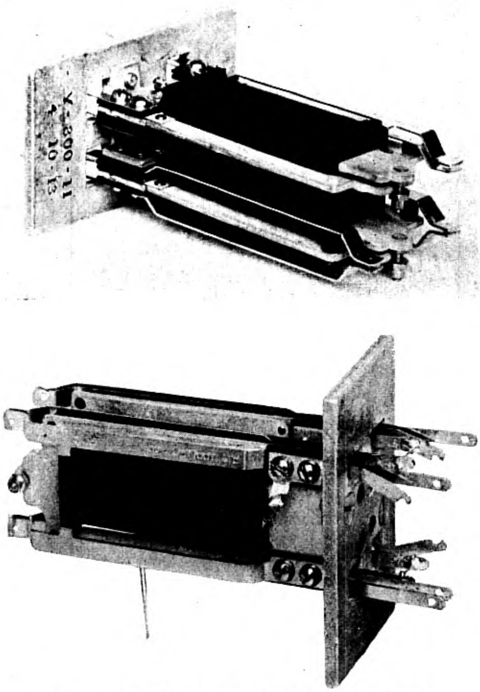


Fig. 10—First punched frame relay

if the design were successful, of employing the same structure with a different winding as a substitute for the cutoff relay. It will be remembered that the line relay had a gravity type armature, whereas the cutoff relay had a reed-hinged armature so the effort to replace two relays of different construction by a single structure was the beginning of an attempt to standardize a type of relay structure which could be used universally.

After some years of development work, a commercial design was completed and punched-type relays were produced as substitutes for the line and cutoff relays. The structures were exactly alike, the relays differing only in windings and in the number of contact-carrying springs with which they were equipped. The development of these relays resulted in a price reduction for the line and cutoff relay unit of about 25 per cent. and a reduction in the mounting space occupied of 40 per cent. The flat core and the manner of suspending the armature on a reed hinge, in order to present the armature to the pole face, were the distinctive features of the new

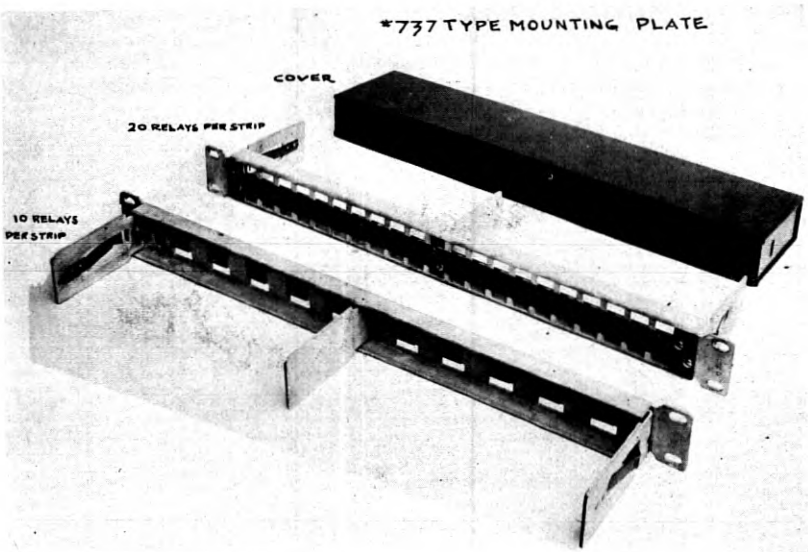


Fig. 11—Mounting plate for strip of punched frame relays

relay structure, as will be seen by referring to Fig. 10, in which the line and cutoff relays may be distinguished because they are equipped with a single pair and a double pair of contacts, respectively. The method of mounting the relays and protecting a strip of 20 with a common dust cover is shown in Fig. 11 from which it will be observed that the mounting plate, all the mounting details and the cover, are products of the punch press.

When it was seen that the development of the new line and cutoff relays was proceeding favorably, development work was also begun on a similar punched-type substitute for the round core supervisory relay which has previously been described. It was known that the quantity of iron in the supervisory relay was greatly in excess of the amount required, as the core flux density was far below saturation when the relay operated over the longer substation loops and the magnetizing ampere turns were reduced by the high resistance of the loop. Advantage was taken of silicon steel, a new material at that time, which had a higher permeability than Norway iron and less pronounced residual magnetic effects, after saturation. In addition, it had greater tensile strength and, since the new type relay core was rectangular in shape and therefore had the stiffness of a beam, it was possible to make a core of silicon steel of such small cross section that the flux density was much higher with a small

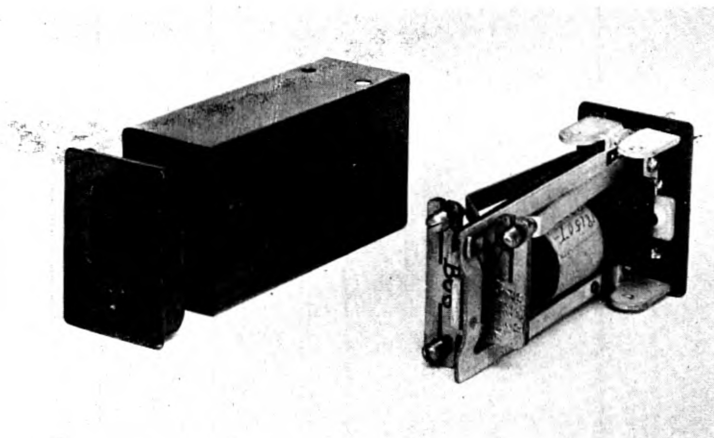
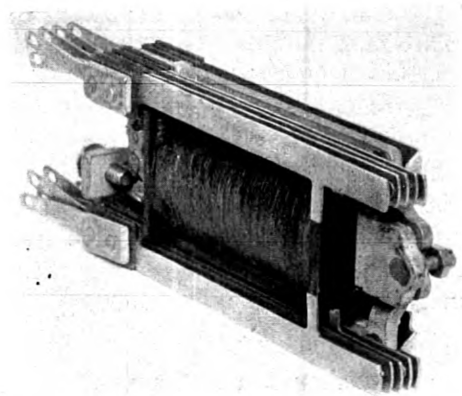


Fig. 12—"B" type relay

magnetizing force than it would be with a Norway iron core of the minimum cross section necessary for structural strength. A supervisory relay was, therefore, produced which was similar in construction to the line and cutoff relays and occupied the same mounting space. It was necessary to develop a dust protecting cover for this new relay which was also cross-talk-proof, in order to prevent the reproduction of conversation by mutual induction between adjacent relays. The design of this relay was such that spring tensions and contact adjustments were controlled by screws mounted in a brass plate at the front of the relay. The increased sensitivity of this relay over that of the round core type permitted the limits for substation loop resistance to be increased from 750 to 1,000 ohms, and the combined resistance of the windings to be reduced from 12 to 9.4 ohms, which decreased the transmission loss in the relay about 30 per cent. In addition, this new relay was superior in flashing ability and also released on a higher number of ampere turns. The mounting space was reduced 25 per cent. Large savings also resulted from a reduction in maintenance costs from approximately $5.00 per switchboard position per year to a negligible amount. The new relay is shown in Fig. 12, which shows the adjusting screws in the plate



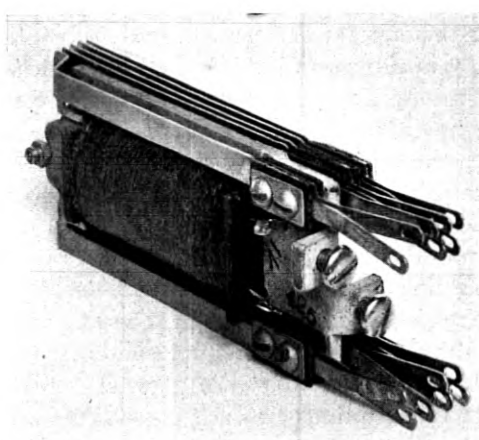


Fig. 13—"E" type relay

in the front of the relay, the cover and the cover cap. By removing the cover cap these screws become accessible and the replacement of the cap does not influence the magnetic conditions or disturb established adjustments.

A General Utility Relay

The success of the punched-type line, cutoff, and supervisory relays suggested the use of this type for a general utility relay which would carry a load of either one pair or several pairs of springs and permit an almost unlimited number of contact spring combinations to be made. This was accomplished by increasing the cross section of the core and armature of the line relay as the increase in iron cross section provided maximum flux with large magnetizing forces. This relay is shown in Fig. 13 and is now manufactured in large quantities with about 3,000 varieties of windings and spring arrangements. About twenty million such relays are already in service and the number is increasing constantly. Had it not been for the development of this punched-type relay, it would have been necessary to greatly increase the manufacturing facilities over those now provided because of the magnitude of the manufacturing operation on the old basis.

Certain Relay Groups

Having outlined the development of the most commonly known relays and given the reasons responsible for major design changes, it will be interesting to consider uses of simple relays in the full mechanical system. In this system the removal of a substation switchhook causes a line relay in the central office to operate and associate a line finder with the calling line, after which a cutoff relay removes the line relay from the circuit as is done in manual practice. A sender is associated with the calling line and the circuit is completed through the substation set dial and a relay in the sender, known as the pulse relay, because it reproduces the dial pulses.

A schematic for illustrating the principle of this circuit is shown in Fig. 14. Referring to this figure, it will be seen that the operation of the pulse relay provides a ground for a slow release relay which in turn extends the circuit of the stepping switch to the back contact of the pulse relay. Suppose that the digit *O* is dialed. Then the resulting current interruptions consist, as shown in Fig. 14, of ten break periods and ten make periods, the final make period being permanent and the remaining nine consisting of approximately one-third of the total time of a single pulse. The first break of the dial opens the circuit through the pulse relay, which releases and opens the circuit of the slow-release relay, but the latter remains operated throughout the break period. The pulse relay when released, provides a ground from its back contact, for the magnet of the stepping switch, through the make contact of the slow release relay. The stepping switch magnet operates the switch armature and holds it in a position to advance the switch a single step when the magnet is released. When the dial contacts close the circuit again the pulse relay re-operates, releasing the stepping switch, which advances one step, and reestablishing the circuit for the slow-release relay. This cycle is repeated for each break and make pulse period in order to advance the stepping switch over the number of terminals corresponding to the digit dialed.

The adjustment of substation dials is such that pulses are sent at a rate of speed of not less than eight, or more than twelve pulses per second. The break period of individual pulses may vary from .045

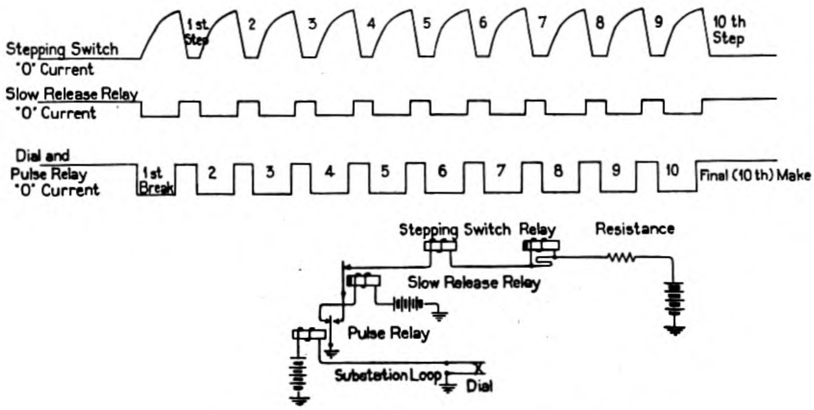


Fig. 14—Curve showing pulsing impulses

to .100 second and the make period may vary from .025 to .050 second. The magnet of the stepping switch must, therefore, complete the movement of the armature in a minimum of .045 second and the switch must advance a single step in a minimum of .025 second. In addition, the slow-release relay must remain operated for a maximum of .100 second; for if it releases during the break-pulse period, the circuit to the stepping switch will be opened. These time values assume that the pulse relay accurately reproduces the dial pulses and it is evident that to accomplish this, its time of operation and release must be independent of the battery potential, between the voltage limits prescribed for the battery, and must also be independent of the differences between the electrical constants of different lengths of substation loops. These are difficult requirements and a punched-type general utility relay, shown in Fig. 13, was used for the purpose as it appeared to be the most suitable available relay. Its time constant, however, is influenced by the electrical constants of the loop with which it is associated; so that the length of the loop effects the speed at which the armature operates and releases and thus causes the relay to introduce some pulse distortion.

An Accurately Adjustable Flat Type Relay

In order to decrease this distortion a new punched-type relay was designed which reproduces dial pulses with much greater accuracy. It will be seen from the picture of this relay shown in Fig. 15 that the armature is light, that the air gaps can be adjusted closely and with great precision, and that the reduction in the inertia of the armature was obtained by changing the position of the supporting

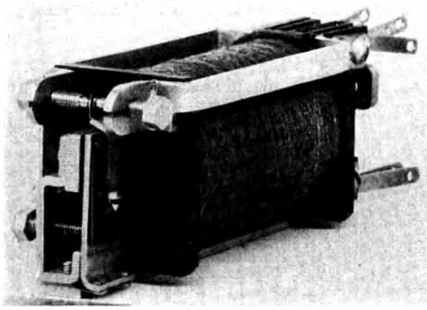


Fig. 15—"L" type pulsing relay

reed hinge. The core of this relay is of small cross section, so that a condition of magnetic saturation is obtained with small current values. With maximum flux on long loops, the increase in current as the length of the substation loop decreases produces very little change in the total flux. Also, changes in the armature air gap as the armature approaches the core do not reduce the reluctance of the magnetic circuit appreciably; so that the armature operates and releases with little time variation irrespective of changes in the electrical constants of the loop.

The slow-release relay, in Fig. 14, is a round-core relay with a reed hinge armature, similar in general construction to the cutoff relay previously described in connection with the early manual system. It is provided with a copper sleeve on the core which acts as a short circuited secondary transformer winding of very low resistance.

Relays in Fundamental Selecting Circuit

For another interesting example of the importance of relay operation in machine switching circuits assume that it is desired to select the fourth terminal in a particular group of a final selector bank as this terminal represents a subscriber's line which has been called from another station. A schematic illustrating the principle of the fundamental circuit for selecting this terminal is shown in Fig. 16. The calling subscriber, by dialing the number of the called station, has

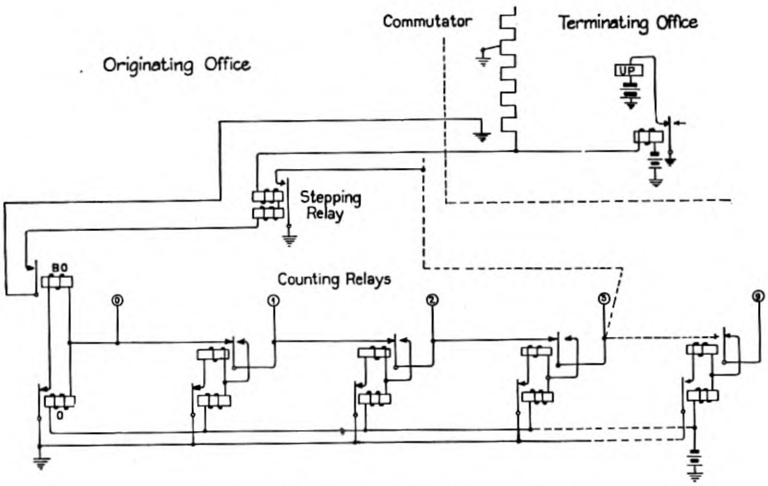


Fig. 16—Schematic of selecting circuit

established the circuit condition shown in this figure through the medium of the sender, so that both the line relay on the final frame in the called office and the stepping relay in the calling office are operated. The circuit is also closed through the up-drive magnet of the final frame, and the selector multiple brush is advancing toward the bank terminal to be chosen. As the selector is driven upward, the commutator brush making contact with the first commutator segment, of the particular group desired, places ground on the inter-office trunk in the called office which shunts down the stepping relay in the calling office. This releases the stepping relay, which had established a circuit when operated through the lower relay of the fourth pair of counting relays and had shunted the upper relay of the pair so it would not operate. The release of the stepping relay removed this shunt and permitted the upper relay to operate, locking both relays through the make contact of the lower relay and transferring the start lead to the lower relay of the third group which operates when the start lead is again grounded through the make contact of the stepping relay. This cycle is repeated for each segment which the commutator brush passes over until the upper relay of the fourth or zero group of counting relays operates and opens the fundamental selecting circuit, thus allowing the line relay in the final frame to release when the commutator brush again removes the shunt. The line relay, on releasing, opens the up-drive circuit and the selector stops with the multiple brush resting on the particular terminal desired.

There are three different types of relays in this circuit. The line relay on the final frame is the general utility punched-type relay of Fig. 13 with the contact spring assembly and mechanical adjustments required by the specific circuit condition. It is evident that this relay must release quickly enough to enable the up-drive clutch magnet to release before the selector is driven beyond the desired terminal or a false bank terminal selection will be made. An examination of some of the factors influencing the release time of the line relay will therefore be of interest.

When the commutator brush made contact with the commutator segment both ends of the inter-office trunk were grounded but before the brush left this segment the condenser charge on the trunk leads was dissipated and the distant end of the trunk was opened by the operation of the upper counting relay of the zero group. On leaving the fourth commutator segment the brush opened the circuit of the line relay which could not release instantaneously because of its own time constant, the transient current through its windings for charging the trunk capacity, and the leak current in its windings resulting from trunk leakage.

The time constant is determined by the electrical and magnetic constants of the relay and for a given winding is inherent to its structure. If the time constant is such that adjustments, for armature air gap, spring tension and contact separation, cannot be made which will enable a relay to meet all the circuit requirements, a different type of relay structure having a more favorable time constant must be used.

The magnitude of the charging current for the trunk is determined by the trunk capacity and is in direct proportion to the length of the trunk which is limited to 12 miles corresponding to a maximum capacity of about 0.84 mf. The limiting open circuit resistance of the trunk is 30,000 ohms and the standard of maintenance is such that the insulation resistance is not allowed to drop below this value.

In addition the maximum resistance of the trunk is 1300 ohms. The line relay must therefore be adjusted to operate over this resistance and in series with the stepping relay when the battery potential is a minimum of 44 volts. It must also be adjusted to release quickly enough to insure the positive selection of a particular terminal when the battery potential is a maximum of 52 volts and both the trunk capacity and trunk leakage are maximum. These are very severe requirements to be met by a relay which is produced commercially in large quantities at a small cost; and more severe conditions such as would result, for example, from increasing the length of the trunk could not be imposed on this particular relay unless the iron structure were made from some new material having more favorable magnetic constants.

The requirements for the stepping relay, however, are more severe than those for the line relay, for the stepping relay must continually operate and release as the commutator brush alternately grounds and frees the trunk in the distant office. Also the insulation resistance and capacity of the trunk exert a somewhat different influence on the functioning of the stepping relay than on the functioning of the line relay. The trunk leakage current resulting from low insulation resistance interferes with the operation of the stepping relay, instead of its release, so it must be adjusted to operate on a minimum battery potential of 44 volts and a minimum trunk insulation resistance of 30,000 ohms. The trunk capacity interferes more seriously with the release of the stepping relay than with the release of the line relay. When the ground is removed from the latter the trunk is at zero potential and the charging current through the relay windings is maintained for a very brief period of time but when the incoming end of the trunk is grounded to release the stepping relay in the distant office, the trunk capacity is fully charged and the discharging current is sustained for a much longer time interval.

The Stepping Relay

The time constant of the line relay is such that it cannot be given adjustments which will enable it to meet the more severe requirements of the stepping relay, and consequently an entirely different type of structure, as shown in Fig. 17, is used for a stepping relay. This design is of particular interest because it is not used for any other purpose and is the only relay of its type in the telephone plant. Many attempts have been made to replace it with some sort of punched type structure that is more adaptable to the established manufacturing methods but they have been ineffectual as yet, for the equivalent combination of sensitivity and reliability and a delicate means of adjustment is difficult to attain. In order to satisfy the severe circuit conditions the stepping relay is adjusted to operate on 10 mil-amperes and not to operate on 9 mil-amperes, a difference of only 10 per cent. in the operate and non-operate adjustments.

The stepping relay must reproduce the pulsations of current originated by the commutator brush with sufficient accuracy to insure

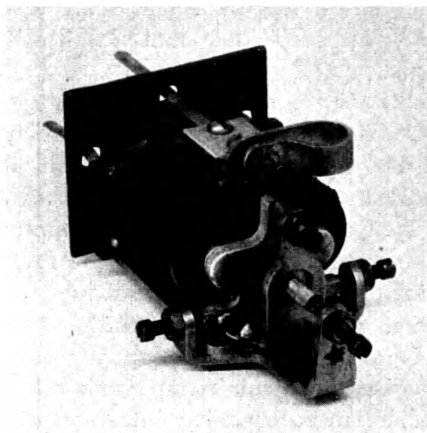


Fig. 17—Stepping relay

the positive operation of the counting relays, for any failure of the latter will result in false selection. The stepping relay must therefore maintain a circuit through its make contact for a sufficient time to enable the lower relay of any counting pair to operate and must open the circuit through the same contact long enough to permit the upper relay of the pair to lock up in series with the lower relay. Since the stepping relay does not always reproduce the commutator pulses perfectly and since any pulse distortion must necessarily reduce the operating time margin for one of the relays of a counting pair, it is evident that rapid operation and reliability of operation are essential characteristics for the counting relays. A punched type relay similar to the line relay cannot operate with sufficient speed. The stepping relay would qualify for speed, but a complete set would require considerable space and would be inconvenient to mount.

The Counting Relay

The relay designed for a counting relay is shown in Fig. 18 and has the qualities of speed and reliability that are required. It is equipped with a light armature, on a pivot suspension, that operates

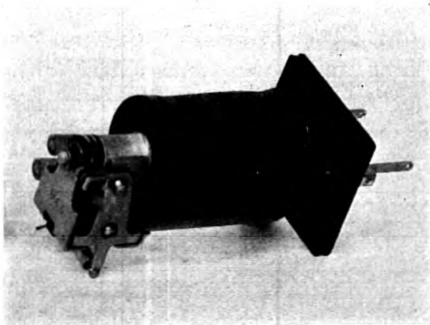


Fig. 18—Counting relay

through a small air gap. The contacts are mounted on rigid springs that cannot be adjusted readily, but which maintain a given adjustment, without change, for a long time. This relay, like the stepping

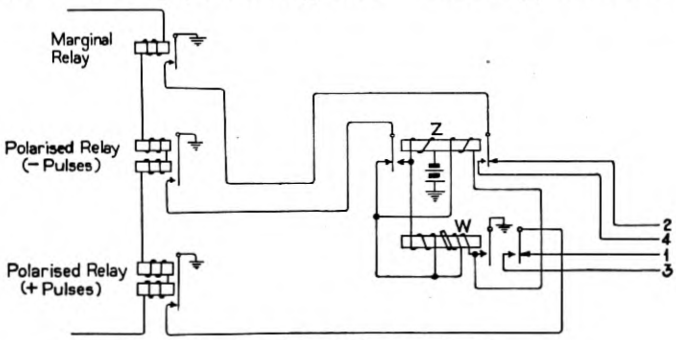


Fig. 19—Call indicator circuit

relay, is unique, in that, it is not used for any other purpose in the telephone system, and in addition all attempts to design a punched type relay that is a satisfactory substitute have, so far, been unsuccessful.

Certain Marginal and Polarized Relays

Another interesting and unusual use of relays is the arrangement at the terminating end of a call indicator trunk from a full mechanical to a manual office. This arrangement consists of three relays in series in the manual office, as shown in Fig. 19. One of them is a marginal relay adjusted to operate on any current greater than a particular value. The other two are polarized relays, one being adjusted to respond to negative pulses only, while the other responds only to positive pulses.

Each digit of any number transmitted over the trunk to the manual office consists of four pulses. The second and fourth of these pulses are always negative, but either or both of them may be a light or heavy negative. The first and third pulses may either be positive or zero, a zero pulse representing a no current interval. This combination of pulses is shown in the following table:

1	2	3	4
+	–	+	–
0	–	0	–

As each pulse interval may consist of either of two kinds of pulses, there are sixteen combinations which can be transmitted but six of them are not used, as only ten are required.

The marginal relay is adjusted to operate on heavy pulses only and as all the positive pulses are light, it does not respond to any positive pulses or the light negative pulses. The negative polarized relay responds to both light and heavy negative pulses and the positive polarized relay responds to all positive pulses. During a zero pulse period all of the relays remain unoperated. For the second and fourth pulse periods the negative polarized relay will be operated and the marginal relay may or may not be operated. During the first and third pulse periods the positive polarized relay may be operated or all the relays may remain unoperated. From the operation of these relays an arrangement of register relays is set up which lights before the manual operator the lamps corresponding to the digits transmitted. The marginal relay used is a counting relay of the type shown in Fig. 18, as this relay has the qualities of sensitiveness, stability and permanence of adjustment that are essential for satisfactory operation. The other two relays are very sensitive polarized relays with micrometer adjustment screws and are representative of the best standards of design for relays of their type. This type of relay is shown in Fig. 20.

The Sequence Switch

Most of the relays previously described were designed to meet specific requirements of unusual severity which limited the design to individual structures having their armatures in close association with the contact carrying springs. Many of the switching operations required for relaying a circuit from point to point through an office can be performed under conditions allowing greater latitude in relay design which has led to the development of several interesting and unusual forms of multi-contact relays in which the armatures

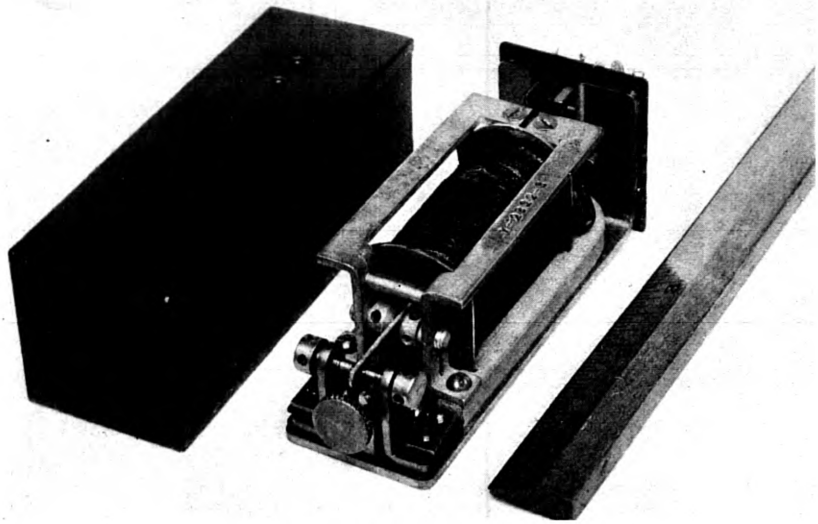


Fig. 20—Call indicator polar relay

indirectly control groups of contact carrying springs. In the development of the machine switching system the work of establishing circuits performed by human relays was transferred to mechanical relays and it soon became evident that the number of individual relay structures of the conventional type required for such a substitution would be so great and the circuit arrangements would be so complicated that the cost would be prohibitive.

The 24 cam sequence switch shown in Fig. 21 is an interesting example of the remote contact control multi-contact relay that not only performs the functions of a multitude of individual relays but actually replaces entire circuits which would require large numbers of relays to control the particular relays that transferred the circuit from point to point. The relay sequence switch shown in the figure is assembled with a shaft that may be rotated into any one of 18 positions which are stamped on an index wheel and are indicated by the position of the wheel with reference to a pointer fixed to the frame of the switch. Each of the circuit switching cams is associated with four brushes and it is possible to so arrange the contact carrying segments on these cams that 6624 different circuit combinations can be established by advancing the switch successively into each of the 18 positions.

The switch is propelled by a driving disc mounted on a power driven shaft that revolves constantly at a speed of 36 r.p.m. The driven disc on the switch in association with the driving disc constitutes a

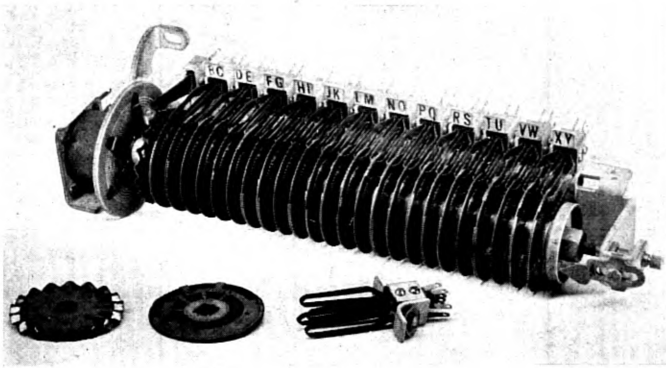


Fig. 21—24 cam sequence switch

friction clutch under the control of an electro-magnet which deflects the driven disc to bring it into relation with the driving disc when it is desired to advance the switch. The electro-magnet corresponds to the winding of an individual relay structure and the driven disc is the armature, the combination of the winding and armature simply serving as a means for controlling the contact relations of a multiplicity of springs.

The Power Driven Selector

The power driven selector shown in Fig. 22 is another example of an entirely different form of multi-contact relay for transferring the three contacts of any one of 500 circuits to the contact springs of a brush that will relay that circuit to any desired point. These 500 circuits are assembled in five groups of 100 each in five banks that are mounted on a frame as shown in the figure. Five brushes, one for each bank, are assembled on a vertical rod in such relation to the banks that the mechanical tripping or release, of any brush brings its springs in contact with the terminals of the bank with which it is associated. The corresponding springs of each of the five brushes are connected in multiple so that in relaying a circuit it is necessary to trip only that brush which is presented to the bank in which the terminals appear. Bringing the brush springs in contact with a par-

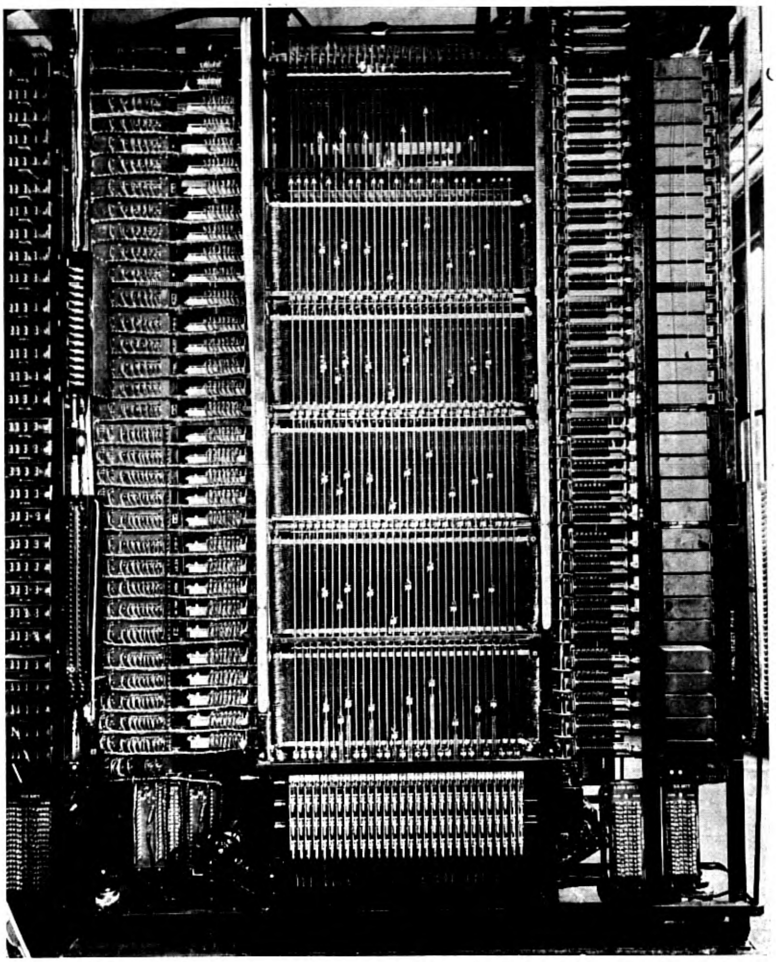


Fig. 22—Power driven selector

ticular group of terminals is referred to as a process of selection and is accomplished by driving the brush rod upward until the desired terminals are reached. When the circuit arrangement is no longer desired the brush rod is driven downward to a normal position where the tripped brush is also restored mechanically to its original condition.

The power for elevating and restoring the brush rod is provided by continuously revolving motor driven steel rolls covered with cork and mounted at the base of the frame. The driving of the brush rod, the tripping of the desired brush, the stopping of the rod and its restoration to normal are all controlled by a series of electro-magnets assembled in a single structure called a clutch which is also mounted at the base of the frame directly in front of the rolls. When a brush rod is driven either up or down, a clutch armature establishes a friction contact between a flat strip of phosphor bronze fastened to the lower end of the brush rod and the cork on the revolving rolls. This clutch is comparable to an individual relay structure with a multiplicity of windings and armatures that are so related that each armature will operate only when its associated winding is energized. The clutch thus does the work of either an exceedingly intricate individual relay or a whole group of less complicated relays. The clutch windings are in effect, relay windings that control the positions of remote contact springs through the operation of armatures which associate or disassociate electro-magnetic and mechanical energy as is desired.

The Step-by-Step Selector

Another type of multi-contact relay in general use that differs in form from both the sequence switch and the power driven selector is the step-by-step selector shown in Fig. 23. It consists of six semicircular contact levels assembled in a bank and an electro-magnet which drives a set of six, double ended, rotary brushes over the terminal arc by means of a driving pawl and ratchet wheel. Each time the magnet is energized and released the driving pawl engages the next tooth on the ratchet wheel which rotates to advance the brushes a single step so that they make contact with the next set of terminals. In 44 successive steps the six brushes move through a complete revolution but as they are double-ended all the possible circuit combinations are set up in the first 22 steps and are then repeated.

In this selector the winding of the electro-magnet corresponds to the winding of an individual relay. The armature in operating elongates a spring that is shown in Fig. 23 and the energy stored in this spring restores the armature to normal and advances the six contact making brushes to the next set of contact terminals. Thus the relay winding and armature control the position of the contact springs through the agency of a flexible mechanical link. The relay winding may be alternately energized and released by current interruptions from an outside source or the armature may be arranged to interrupt the circuit through the winding by opening a pair of contacts in the operated position to advance the selector by self interruptions.

Relays in Toll Circuits

Supervision on all of the longer toll circuits and on most of the shorter ones is provided on what is known as a ringdown basis. This

Fig. 23—Step-by-step rotary switch

usually involves a ringup relay at each end of the line, which operates in response to 20-cycle signaling impulses. These impulses may be transmitted over the line from one office to the other or they may originate in the same office as the relay and be impressed on the line by the operation of a so-called composite ringer in response to signals of a different frequency. The ringup or drop relay provides the signal in the toll switchboard. It is usually removed from the circuit when the line is taken up by the operator and the supervision is then transferred to the toll cord.

The toll cord supervisory circuit is shown in Fig. 24 and illustrates a typical condition which has imposed particular requirements on the relays involved. The signal receiving relay *A* may be bridged directly across the line conductors, in which case its winding must be of such high impedance that it does not materially affect the efficiency of the talking circuit. Experience has indicated that with the windings commonly used on relays there is some chance of sufficient short circuited turns to materially reduce the inductance of the winding. This may occur in a relay which would otherwise give satisfactory operation, the short circuited turns merely reducing

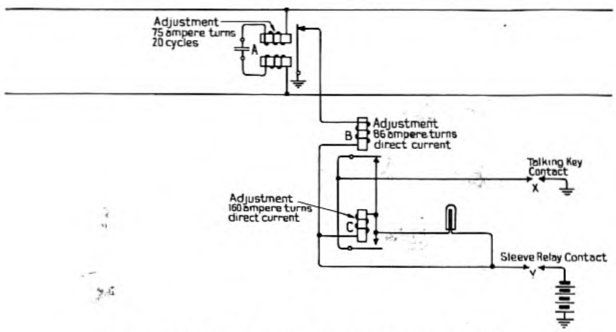


Fig. 24—Toll cord supervisory circuit

slightly the low frequency or direct-current efficiency. For this reason, the relay winding has been divided into two parts, on separate cores, either one of which has sufficient inductance to safeguard the telephone transmission.

The incoming 20-cycle signaling current may be of small value and the portion through the relatively high impedance of the relay will be still smaller so that this relay must be extremely sensitive. The relay has small moving parts and a comparatively light spring tension. These factors contribute to sensitive operation but also permit the opening of the contact on impulses other than those intended for signaling. Such impulses are usually of short duration and the other relays of the circuit have been designed to limit their effect to prevent false signals.

Both relays "*B*" and "*C*" are of the same type, designed to operate with a slight time lag so that other things being equal they would be expected to operate at the same time when the circuit is closed at *x* and *y*. Relay "*B*," however, receives, under the worst condition, 150 per cent. of its rated operating current, while relay "*C*" receives 105 per cent. This will tend to make relay "B" quicker in operation than relay "*C*," so that when the battery and ground are connected to the circuit, relay "*B*" will operate first and open the winding of relay "*C*." This is therefore the normal condition of the circuit and is further insured by the fact that the opening of the winding of "*C*" occurs at a back contact of relay "*B*" while the locking of "*C*" occurs only after the relay has pulled up to close its front contact.

The sequence of operation and release resulting from this series of relay operations affords protection against false signals since relay "*A*" must operate continuously until "*B*" has released and "*C*" has operated before the lamp circuit is closed. Relay "*B*," in addition to being slow in operation, is also slow to release, so that the time interval thus introduced tends to bridge over any transient impulses that may tend to operate the signal.

The slow operation of relays is secured by means of a copper sleeve over the relay core. Slow operation results from the transient condition existing during the time between the application of voltage to the relay winding and the building up of the magnetic field to a steady state. Slow release results from the transient condition existing during the time between the removal of the voltage from the relay winding and the decay of the magnetic field until the magneto-motive force falls below the armature restoring force. These conditions are more easily seen when the relay winding is considered as the primary of a transformer and the copper sleeve as a short-circuited secondary winding consisting of a single turn having a very low resistance. The operating current, before it reaches its steady value, may be considered as an alternating current of one-quarter of a cycle, starting from zero and building up to a maximum value. Slow operation of the armature results from opposing the building up of the flux in the core. Slow release is due to retarding the decay of the flux in the core. The speed at which the armature operates or releases is not changed but in the first case the application of the magneto-motive force required to move the armature is delayed, and in the second case the removal of the magneto-motive force holding the armature in the operated position is also delayed. When a voltage is first applied to the terminals of the winding, the current tends to build up and establish the magnetic flux at its maximum value in the relay core. The instant the flux threads the copper sleeve, a voltage is induced in the latter, causing a current to flow in it. This current in the copper sleeve sets up a flux in the same magnetic path which opposes the flux building up from the current in the relay winding. Due to leakage, the winding flux is greater than the opposing flux set up by the sleeve and the resultant flux continues to build up until it reaches a maximum value. This opposition of the winding flux and the flux produced by the induced current in the copper sleeve increases the time for the building up of the magnetic force necessary to move the armature from the normal position. It also increases the time for such a reduction in the magnetic force as will permit the armature to release.

The slow release feature is further secured by omitting the stop pins which are usually provided between the armature and the pole

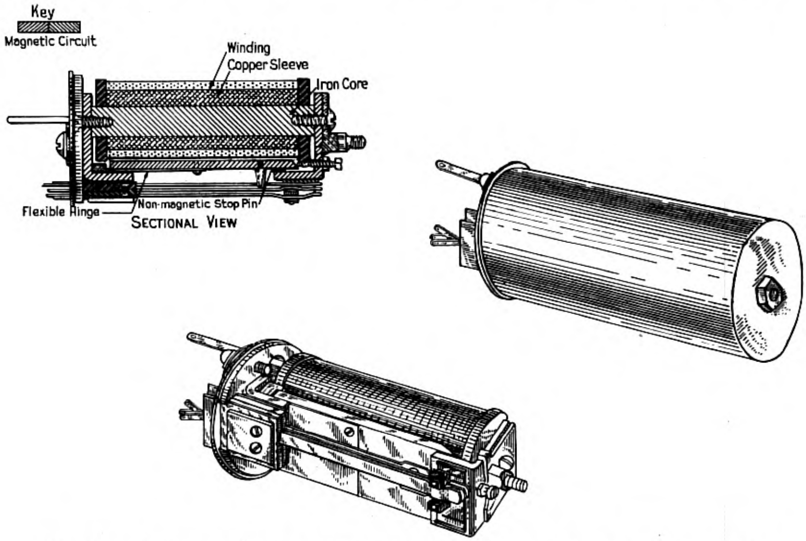


Fig. 25—Sectional view of No. 162 type relay, a slow operating relay

piece. This tends to delay the decline in flux through the magnetic circuit when the current is interrupted. Fig. 25 illustrates diagrammatically the structure of these relays.

RELAYS OF THE COMPOSITE RINGER

A somewhat similar use of relays is to be seen in the composite ringer circuit mentioned above. The relay circuit of such a ringer is shown in Fig. 26, in simplified form. This circuit is designed to receive 20-cycle signals from the switchboard and transmit out on the line signals of a higher frequency and to receive the higher frequency impulses and in turn transmit 20 cycles to the switchboard. In this case, the 20-cycle relay "*A*" does not meet the requirement for high impedance since protection to the telephone circuit is afforded by coil "*C*." A single core is therefore satisfactory and a positive make-contact relay is used. In this case, the chief requirement is that relay "*B*" should be slow in operating.

The chain of relays operating from the high frequency signals consists of relays *D*, *E* and *F*. Relay "*D*" must be a very sensitive structure in this case and a polarized relay with a vibrating contact has commonly been used. The circuit requirements are such that the energy available for the operation of this relay is seldom more than a few hundred microwatts and may be much less. The cir-

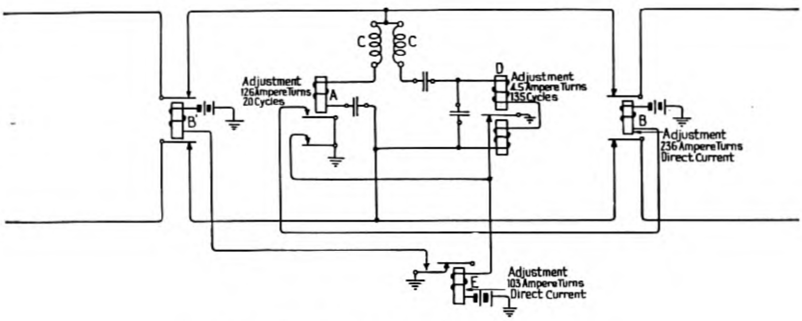


Fig. 26—Composite ringer circuit.

cuits are being designed on the basis of giving reliable operation on 20 microwatts. The operation of relay "*D*" releases relay "*E*" which in turn operates relay *F*.

Where such a circuit depends on the operation of a vibrating contact relay, the current through this contact is of vital importance. Whenever the contact is closed, current tends to flow through the winding of relay "*E*." Fig. 27 illustrates the effect of very weak signaling currents and of currents sufficient to give proper operation. The current values through the vibrating relay winding and through the winding of the secondary relay are shown for two different typical conditions. Also, the average or effective value in winding "*E*" is shown.

A circuit feature which has recently been introduced to increase the sensitivity of relay "*D*" and to improve the operation of the secondary relay consists in the introduction of a condenser and the operation of the vibrating contact as a normally open contact. The closing of the contact charges a condenser which tends to operate the secondary relay by its discharge as soon as the contact opens. By this combination, the effect noted in Fig. 27 is eliminated and positive operation of the secondary relay is secured as soon as the armature vibrates sufficiently to make contact. The local circuit embodying this feature is shown in Fig. 28. Referring to this figure and to Fig. 26, relays "*D*" and "*E*" represent the alternating current relay and the secondary relay in each case. In the one case, however, relay E' operating when relay "D" operates gives positive release of relay "E" instead of introducing an uncertain resistance in its circuit.

This circuit embodies several features which are not common in relay systems. The operation of relay E' is dependent on the values

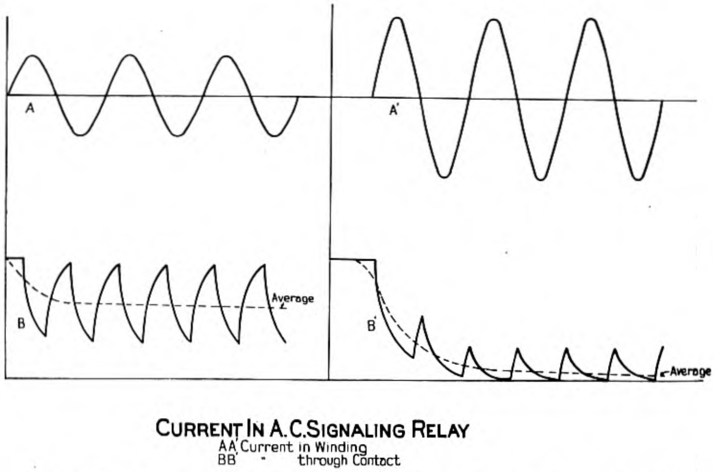


Fig. 27—Curve showing signal impulses in a.c. signaling relay
AA[1], current in winding
BB[1], current through contact

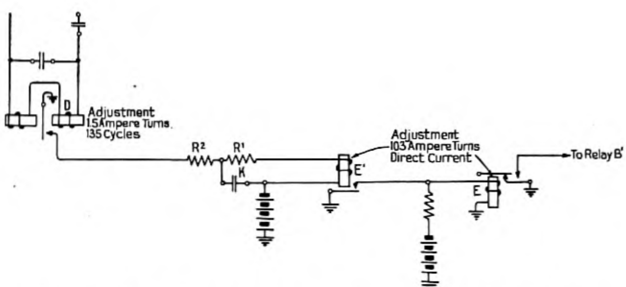


Fig. 28—Circuit for a make contact 135 cycle relay

selected for resistances R' and R^2 and for condenser K. These values must be such that the current in the relay winding is maintained during the opening of the contact of relay "D" by means of the discharge current from the condenser. On the other hand K and R^2 must be proportioned to limit the arcing of the contact of "D" at the frequency of the signaling impulses. The method of releasing relay "E" by short-circuiting its winding has advantages over opening the circuit for the purpose under consideration. The arcing at the contact of relay E' is less severe than would be the case if an inductive circuit were broken.

An added feature which has been incorporated in the mechanical design of relay "D" and which has an important bearing on its performance electrically, is an adjustment limiting the armature travel. This limitation of movement prevents a wide deflection when the relay receives excessive current. Such deflection would tend to set the armature into vibration and would result in a sufficient number of impulses to operate relay E' and cause false signals.

The Vacuum Tube

The vacuum tube is used for the relaying of energy in a number of ways. It may be connected in circuit to amplify the received impulses in which case it sends out energy from a local source with the same wave shape as that of the received current. In this case the tube serves to relay the impulses with as little distortion as possible. In the case of a tube used as a modulator or a demodulator it is required to combine or separate impulses of different character, the two operating together to preserve the same impulses at the output of the demodulator tube as is received at the input of the modulator. The impulses which are transmitted between the two tubes have an entirely different wave form and may be amplified any number of times by means of amplifier tubes without affecting the action of the modulator and demodulator.

The vacuum tube may also be used as a rectifier to convert alternating current to direct or pulsating current or it may be used as an oscillator to produce alternating current from a local source of direct current. In all of these applications of vacuum tubes, the tubes serve as relays to introduce a fresh supply of energy or a desired wave form or a combination of the two to serve their purposes in the communication system.

Relays for Telegraph Circuits

The use of relays for telegraph circuits presents an entirely different set of problems than those usually encountered in the consideration of telephone circuits. Most telegraph relays are used for repeating signals from one circuit to another rather than for switching local circuits. While some marginal operating conditions are imposed on telephone relays there is not the wide range of operating conditions to be met under which most telegraph relays are required to operate. The numbers involved are usually much less so that economies in production play a somewhat less important role and the cost is not quite such an important item. Similarly the methods of assembly and mounting afford a somewhat wider latitude than can be permitted where many thousands of relays must be mounted in comparatively small space.

Because of the exacting requirements imposed on telegraph relays and to insure continuity of service as far as possible, they are usually made interchangeable to a much greater degree than telephone relays. They may be connected by means of screws instead of soldered con nections or they may be inserted in the circuit by means of spring clips in a connecting block.

In a telegraph system speed of operation and reliability are the most important requirements and are very large factors in determining the mechanical design of the relays. The relay must operate quickly and accurately so as to cause as little distortion as possible to the signals. In addition it must be extremely rugged and maintain its adjustment well throughout long continued operation. A very ordinary day's work for a telegraph relay requires the reliable operation of its contacts several hundred thousand times and it may be called upon to open and close its contacts a million times a day. Where a telephone relay might hesitate and still pull up and perform its function properly or might make uncertain contact at first, such behavior on the part of a telegraph relay would result in false impulses and would quickly call for a readjustment or a change of relays.

With the exception of some of the alternating current signaling relays in telephone circuits the energy available in telegraph relays is usually less than that available for telephone relays. The more sensitive relays are called upon to operate from line current which has been attenuated by leakage or by parallel paths and which may have been limited at the distant station. Systems operating over open wire lines are usually restricted to about .075 ampere at the sending end and in cable the normal current is about .005 ampere. This difference is not as great in actual operation as would appear since the open wire system operates on a ground to ground basis and the cable on a metallic basis. In operating from ground at one station to ground at another differences in ground potential and leakages occur which require a greater margin than is necessary with the metallic system.

If satisfactory telegraph service is to be rendered, particularly on long circuits involving a number of relaying points, it is essential that the telegraph relays employed have such operating characteristics that they introduce as little distortion to the signals as possible. It has been found that the polarized type of relay fulfills this condition to a greater extent than the neutral type of relay which is used in local circuits and in some telegraph circuits where extreme accuracy is not required. The polar relay permits arrangements of circuits which minimize the effect of poor wave shape and line leakage. It also is more easily adapted to variations in current strength and may be adjusted to give more accurate repetition of the signals under all conditions.

A number of important developments in telegraph relays have led up to the relay shown in Fig. 29. This relay gives reliable operation with 4-ampere turns in the winding and by careful adjustment may be made to operate on a small fraction of that.

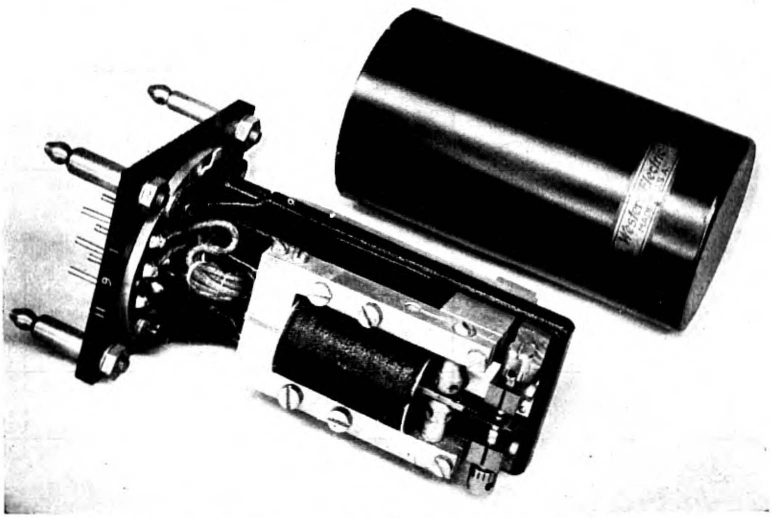


Fig. 29—Photograph of telegraph type relay

While the telegraph relay may be called upon to operate on very small energy, its contact must be capable of handling much larger quantities. Due to the speed of operation desired and to the dependence on accurate transmittal of each impulse the contacts must operate without chatter or vibration. Great care has been taken in the design of the relays and the circuits to protect the contacts and insure good operation. Chatter can be largely eliminated by careful mechanical design and the effect of the arc set up when the contact is called upon to break the current in a circuit carrying several watts can be minimized somewhat by means of, so-called, spark killers. These consist of condensers and resistances so proportioned as to absorb the force of the arc in the charging of the condenser when the contact opens. They can be utilized still further in modifying the shape of the transmitted wave by the charging of the condenser when the contact is opened.

Telegraph relays and their applications have been referred to in this paper only in the most general terms because of the variety of their forms and uses. It is planned to cover this as well as other subjects pertaining to relays in a series of later papers.

Some Applications of Statistical Methods to the Analysis of Physical and Engineering Data

By W. A. SHEWHART

SYNOPSIS: Whenever we measure any physical quantity we customarily obtain as many different values as there are observations. From a consideration of these measurements we must determine the *most probable value;* we must find out *how much* an observation may be expected to vary from this most probable value; and we must learn as much as possible of the *reasons why* it varies in the particular way that it does. In other words, the real value of physical measurements lies in the fact that from them it is possible to determine something of the nature of the results to be expected if the series of observations is repeated. The best use can be made of the data if we can find from them the most probable frequency or occurrence of any observed magnitude of the physical quantity or, in other words, the most probable law of distribution.

It is customary practice in connection with physical and engineering measurements to assume that the arithmetic mean of the observations is the most probable value and that the frequency of occurrence of deviations from this mean is in accord with the Gaussian or normal law of error which lies at the foundation of the theory of errors. In most of those cases where the observed distributions of deviations have been compared with the theoretical ones based on the assumption of this law, it has been found highly improbable that the groups of observations could have arisen from systems of causes consistent with the normal law. Furthermore, even upon an a priori basis the normal law is a very limited case of a more generalized one.

Therefore, in order to find the probability of the occurrence of a deviation of a given magnitude, it is necessary in most instances to find the theoretical distribution which is more probable than that given by the normal law. The present paper deals with the application of elementary statistical methods for finding this *best* frequency distribution of the deviations. In other words, the present paper points out some of the limitations of the theory of errors, based upon the normal law, in the analysis of physical and engineering data; it suggests methods for overcoming these difficulties by basing the analysis upon a more generalized law of error; it reviews the methods for finding the best theoretical distribution and closes with a discussion of the magnitude of the advantages to be gained by either the physicist or the engineer from an application of the methods reviewed herein.

INTRODUCTION

WE ordinarily think of the physical and engineering sciences as being exact. In a majority of physical measurements this is practically true. It is possible to control the causes of variation so that the resultant deviations of the observations from their arithmetic mean are small in comparison therewith. In the theory of measurements we often refer to the "*true value*" of a physical quantity: observed deviations are considered to be produced by errors existing in the method of making the measurements.

With the introduction of the molecular theory and the theory of quanta, it has been necessary to modify some of our older conceptions. Thus, more and more we are led to consider the problem of measuring any physical quantity as that of establishing its most probable value. We are led to conceive of the physico-chemical laws as a statistical determinism to which "the law of great numbers" [1] imparts the appearance of infinite precision. In order to obtain a more comprehensive understanding of the laws of nature it is becoming more necessary to consider not only the average value but also the variations of the separate observations therefrom. As a result, the application of the theory of probabilities is receiving renewed impetus in the fields of physics and physical chemistry.

Statistical Nature of Certain Physical Problems. As typical of the newer type of physical problem, we may refer to certain data given by Prof. Rutherford and H. Geiger.[2] In this experiment the number of alpha particles striking, within a given interval, a screen subtending a fixed solid angle was counted. Two thousand six hundred and eight observations of this number were made. The first column of Table I records the number of alpha particles striking this screen within a given interval. The second column gives the frequency of occurrence corresponding to the different numbers in the first column.

TABLE I

No. of Alpha Particles	Observed Frequency of Occurrence
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	0
13	1
14	1

It is obviously impossible from the nature of the experiment to attribute the variations in the observed numbers to errors of observation. Instead, the variations are inherent in the statistical nature of the phenomenon under observation.

[1] Each class of event eventually occurs in an apparently definite proportion of cases. The constancy of this proportion increases as the number of cases increases.

[2] *Philosophical Magazine*, October, 1910.

The questions which must be answered from a consideration of these data are typical. For example, we are interested to know how a second series of observations may be expected to differ if the same experiment were repeated. The largest observed frequency corresponds to four alpha particles, although what assurance is there that this is the most probable number? What is the probability that any given number of alpha particles will strike the screen in the same interval of time? Or again, what is the maximum number of alpha particles that may be expected to strike the screen? All of these questions naturally can be answered providing we can determine the most probable frequency distribution.

Statistical Nature of Certain Telephone Problems. The characteristics of some telephone equipment cannot be controlled within narrow limits much better than the distribution of alpha particles could be controlled in the above experiment. We shall confine our attention primarily to a single piece of equipment. The carbon microphone. For many reasons it is necessary to attain a picture of the way in which a microphone operates. It is necessary to find out why carbon is the best known microphonic material. In order to do this we must measure certain physical and chemical characteristics of the carbon and compare these with its microphonic properties when used under commercial conditions. In the second place it becomes necessary to establish methods for inspecting manufactured product in order to take account of any inherent variability, and yet not to overlook any evidence of a "trend" in the process of manufacture toward the production of a poor quality of apparatus. In the third place it so happens that the commercial measure of the degree of control exhibited in the manufacture of the apparatus must be interpreted ultimately in terms of sensation measures given by the human ear. That is, the first phase of the problem is purely physical; the second is one of manufacturing control and inspection and the third involves the study of a variable quantity by means of a method of measurement which in itself introduces large variations in the observations.

In one of the most widely used types of microphones there are approximately 50,000 granules of carbon per instrument. Each of these granules is irregular in contour, porous and of approximately the size of the head of a pin. If such a group of granules is placed in a cylindrical lavite chamber about ½-inch in diameter and closed at either end with gold-plated electrodes; if this chamber is then placed on a suspension free from all building vibrations and carefully insulated from sound disturbances; if automatically controlled mechanical means are provided for rolling this chamber at any desired speed; if all of the air and sorbed gases are removed from the carbon chamber and pure nitrogen is substituted; if the mean temperature is kept constant within 2° C; and if means are provided for measuring the resistance of the granules when at rest by observing the voltage across the two electrodes while current is allowed to flow for a period less than 1/200 of a second, it is found that the resistance (for most samples of carbon) may be determined within a fraction of one per

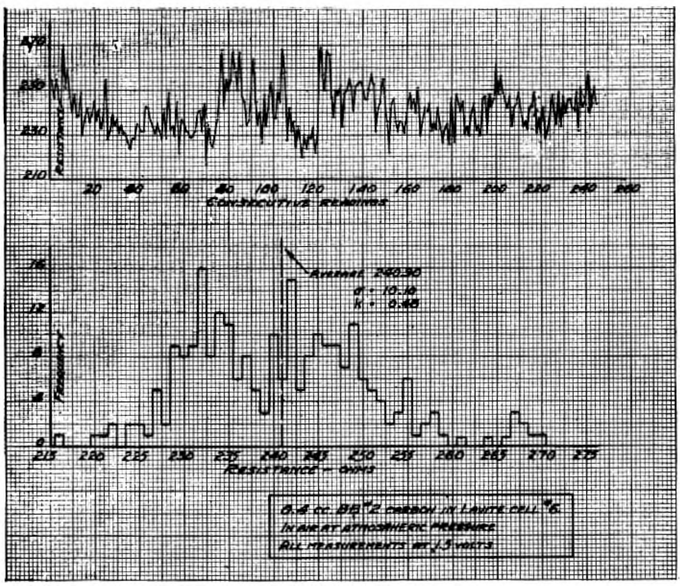


Fig. 1

cent. If, however, the button is rotated (even as slowly as possible) and then brought to rest, the resistance may differ several ohms from its first value. If a large number of observations are made after this fashion, we may expect to find for certain samples of carbon a set of values such as given in Fig. 1. The 270 observations of resistance reproduced in this figure were made on a sample of carbon at 1½ volts under conditions quite similar to those outlined above. The observed variation is from approximately 215 to 270 ohms. The upper curve is that of the resistance vs. the serial number of the readings. There is no apparent trend in the change of resistance from one reading to another. The lower curve in this figure shows the frequency histogram of the results. Attention is directed to the wide variation in the observations, and to the fact that the frequency histogram appears to be bimodal.[3] Methods of dealing with such distributions will be considered.

Samples of carbon having different molecular surface structures have different resistances. To put it in a still more practical way, if the manufacturing process is not controlled within very narrow limits, wide variations are produced in the molecular properties of the carbon. The microphonic properties of these carbons are therefore different. One of the problems with which we have been concerned is to determine the relationship existing between the physical and chemical characteristics of the carbon and the resistance of the material when measured under different conditions. We are obviously dealing in this case with problems involving the measurement of physical quantities which cannot be controlled even in the labora-

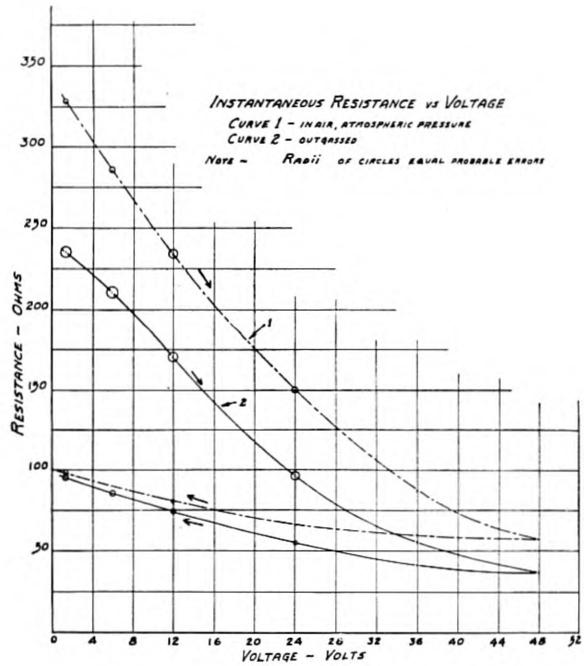


Fig. 2

[3] If curves which touch the axis at $+\infty$ and $-\infty$ have more than one value of the variable for which the derivative of the frequency in respect to the variable is equal to zero—the points being other than that for which the frequency is zero—these curves are referred to as bimodal, trimodal, etc. The modal value is the most probable one and is of particular interest in unimodal curves.

tory. If we remove the air and measure the resistance at different voltages, we may expect to find changes in the resistance similar to those indicated in Fig. 2. Curves 1 and 2 were taken for increasing voltages. The return curves were taken with decreasing voltage. Removal of the air from this particular sample of carbon produces comparatively large changes in the resistance. The resistance at 1½ volts is several times that at 48 volts. These curves were taken under conditions wherein all of the other factors were controlled. A sufficient number of observations was made in each case in order to establish the probable errors of the points as indicated by the radii of

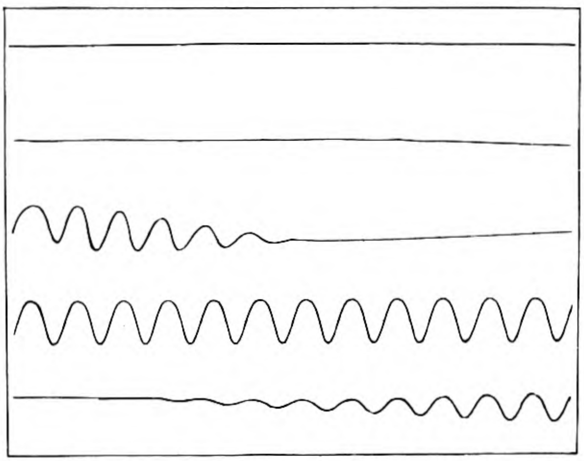


Fig. 3—Possible Types of Breathing of Granular Carbon Microphone.

the circles. If this same experiment were carried on at a different temperature, radically different results would be obtained.

If, instead of allowing the current to flow for a short interval of time, a continuous record is made of the resistance of the carbon while practically constant current flows through the carbon, the resistance will be found to vary. The maximum resistance reached in certain instances may amount to several times the minimum value. In general, this phenomenon is attributed to the effects of gas sorbed on the surface of the material. Transmitters cannot be made of lavite so that the expansions and contractions of the piece parts thereof augment the changes in resistance. This phenomenon, termed "breathing," may be, but seldom is, regular or periodic. An exceptional case of breathing is shown in Fig. 3. This was obtained with a special type of carbon in a commercial structure. The curves themselves represent the current through the transmitter and, therefore, are inversely proportional to the resistance. All five curves were obtained with the same carbon in the same chamber by varying merely the configuration of the granules by slightly tapping the carbon chamber.

All of these effects can be modified to a large extent by varying the process of manufacture of the granular material. In practice it is necessary to know why slight changes in the manufacturing process cause large variations in the resistance characteristics of the carbon. The same process that improves one microphonic property may prove a detriment to another. It is in the solution of some of these problems that statistical methods have been found to be of great value in the interpretation of the results.

Whereas the physicist ordinarily works in the laboratory under controlled conditions, the engineer must work under commercial conditions where it is often impractical to secure the same degree of control. More than 1,500,000 transmitters are manufactured every year by the Bell System. Causes of variation other than those introduced by the carbon help to control the transmitter. For example, variations may be introduced by the process of assembly, or by differences in the piece parts of the assembled instrument. The measure of the faithfulness and efficiency of reproduction depends fundamentally upon the human ear. Obviously all transmitters cannot be tested. Instead, we must choose a number of instruments and from observations made on these determine whether or not there is any trend in the manufactured product. Naturally we may expect to find certain variations in the results according to the rules of chance. To take the simplest illustration, we may flip a coin 6 times. Even if it is symmetrical we may expect occasionally to find all heads and occasionally all tails, although the most probable combination is that of 3 heads and 3 tails. We must, therefore, determine first of all whether or not the observed variations are consistent with those due to sampling according to the laws of chance. If there is an apparent trend in product, the data should be analyzed in order to determine, if possible, whether it is due to lack of control in the manufacture of carbon or to some other set of causes such as mentioned above. Because of economic reasons we must keep the number of observations at a minimum consistent with a satisfactory control of the product. Here again it has been found that the application of statistical methods is necessary to the solution of the problems involved.

Before considering the problem of the measurement of efficiency and quality of the transmitter, let us consider the schematic diagram of the telephone system as shown in Fig. 4. Essentially this consists of the transmitter, the line and the receiver. The oldest method of measurement is to compare one transmitter against a standard in the following way. An observer calls first in the standard and then in the test transmitter, while another observer at the receiving end judges the faithfulness of reproduction. The pressure wave striking the transmitter diaphragm varies with the observer and also with the degree of mechanical coupling between the sound source

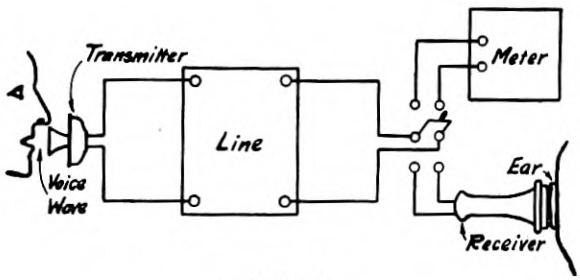


Fig. 4

and the diaphragm of the instrument. The judgment of the observer at the receiving end in influenced by physiological and psychological causes. Obviously it is desirable that such a method be supplanted by a machine test which will eliminate the variabilities in the sound source and in the human ear. Up to the present time the nature of speech and the characteristics of the human ear are not known sufficiently well to establish either an ideal sound source or an electrical meter to replace the human voice and ear respectively. The best that can be done is to approximate this condition. Even though the meter readings may be the same, the simultaneous observations made with the ear in general will be different. A calibration of the machine must, therefore, depend upon a study of the degree of correlation between the average measure given by the machine and that given by the older method of test.

Thus, we see how special problems arise in the fields of both physics and engineering wherein it is impossible to control the variations. In what way, if any, are these problems related, or is it necessary to attack each one in a different manner? We shall see that all of these problems are in a way fundamentally the same and that the same method of solution can be applied to all of them. This is true because it is necessary to determine in every instance the law of distribution of the variable about some mean value.

Why Do We Need to Know the Law of Deviation of the Different Observations About Some Mean Value?

In all of the above problems as in every physical and engineering one, certain typical questions arise which can be answered only if we know the law of distribution $y=f(x)$ of the observations where y represents the frequency of occurrence of the deviations x from some mean value. At least three of these questions are the same for both fields of investigation.[4]

Let us consider the physical problem. From a group of n observations of the magnitude of a physical quantity, we obtain in general n distinct values which can be represented by $X_1, X_2, \ldots X_n$. From a study of these we must answer the following questions:

1. What is the most probable value?
2. What is the frequency of occurrence of values within any two limits?
3. Is the set of observations consistent with the assumption of a random system of causes?

The answers to these questions are necessary for the interpretation of Prof. Rutherford's data referred to above: They are required in order to interpret the data presented in Fig. 1 which are typical of physical and chemical problems arising in carbon study; these same answers are fundamentally required in the analysis of all physical data. These questions can be answered from a study of the frequency distribution. If this be true, it is obvious that the statistical methods of finding the best distribution are of interest to the physicist.

Let us next consider the engineering problem where we shall see that the same questions recur. Assuming that manufacturing methods are established to produce a definite number of instruments within a fixed period, one or more of the characteristics of these instruments must be controlled. We may represent any one of these characteristics by the symbol X. The total number of instruments that will be manufactured is usually very indefinite. It is, however, always finite. Even with extreme care some variations in the methods of manufacture may be expected which will produce

[4] In order to calibrate the machine referred to in a preceding paragraph and also to determine the relationships between the physico-chemical and microphonic properties of carbon, it was necessary to study the correlation between two or more variables, but in each case it was necessary to determine first the law of distribution for each variable in order to interpret the physical significance of the measures of correlation because this depends upon the laws of distribution. The reason for this is not discussed in the present paper, for attention is here confined to the method of establishing the best theoretical frequency distribution derived from a study of the observations.

variations from instrument to instrument in the quantity X. After the manufacturing methods have been established, the first problem is to obtain answers to the following questions:

1. What is the most probable value of X?

2. What is the percentage of instruments having values of X between any two limits?

3. Are the causes controlling the product random, or are thev correlated? [5]

In this practical case we must decide to choose a certain number of instruments in order to obtain the answers to these questions; that is, to obtain the most probable frequency distribution. We must, however, go one step further. We must choose a certain number of instruments at stated periods in order to determine whether or not the product is changing. How big a sample shall we choose in the first place, and how large shall the periodic samples be? Obviously it is of great economic importance to keep the sample number in any case at a minimum required to establish within the required degree of precision the answers to the questions raised.

The close similarity between the physical and engineering problems must be obvious. Naturally, then, we need not confine ourselves in the present discussion to a consideration of only the problems arising in connection with the study of those microphonic properties of carbon which gave rise to the present investigation. Several examples are therefore chosen from fields other than carbon study. However only those points which have been found of practical advantage in connection with the analysis of more than 500,000 observations will be considered.

The type of inspection problem may be illustrated by the data given in Table II.

The symbol X refers to the efficiency of transmitters as determined in the process of inspection: N represents the number of instruments measured in order to obtain the average value X. The first four rows of data represent the results obtained by four inspection groups G_1, G_2, G_3 and G_4. The results given are for the same period of time. The next three rows are those for different machines M_1, M_2 and M_3. The last row gives the results of single tests on 68,502 transmitters, a part of which was measured on each of the three machines. The third column in the table gives the standard deviations. It will be observed

[5] The significance of this question will become more evident in the course of the paper. We shall find that, if the causes are such as to be technically termed random, we can answer all practical questions with a far greater degree of precision than we can if the causes are not random.

TABLE II

INSPECTION DATA ON TRANSMITTERS

	$\overline{X}$	σ	$\frac{3\sigma}{\sqrt{N}}$	N	k	σ_k	β_2	σ_{β_2}	$\sigma_{\overline{X}}$	$3\sigma_k$	$3\sigma_{\beta_2}$	Pearson Type
G_1	.548	.739	.0131	4510	− .214	.056	4.152	.073	.011	.108	.219	IV
G_2	.740	.896	.0533	2540	− .949	.049	4.426	.097	.018	.147	.291	VI
G_3	.766	.762	.0568	1620	− .109	.061	5.176	.122	.019	.183	.366	VI
G_4	.934	.677	.0398	2610	−1.413	.048	7.677	.096	.013	.144	.288	IV
M_1	−1.66	1.32	.0386	10855	− .70	.024	3.128	.047	.013	.072	.141	
M_2	−1.69	1.07	.0300	11577	− .84	.023	4.240	.046	.010	.069	.138	
M_3	−1.79	1.04	.0510	3749	− .56	.040	3.628	.080	.017	.120	.240	
Machines 1, 2, 3	−1.641	1.14	.0131	68502	− .80	.009	Out		.004	.027		I

that comparatively large differences exist between the averages obtained for different groups of transmitters by different groups of observers. Similarly, comparatively large variations exist in these averages even when taken by the machines (the large difference between the sensation and machine measures is due to a difference in the standard used, corrections for which are not made in this table).

Are these differences significant? Is product changing? That is, are the manufacturing methods being adequately controlled? Are these results consistent with a random variation in the causes controlling manufacture? These are the questions that were raised in connection with the interpretation of these data. The ordinary theory of errors gives us the following answer. It will be recalled that the standard deviation (or the root mean square deviation) of the average $\sigma_{\overline{X}}$ is equal to $\frac{\sigma}{\sqrt{N}}$. Also, from the table of the normal probability integral we find that the fractional parts of the area within certain ranges are as follows: For the ranges $\overline{X}\pm\sigma$, $\overline{X}\pm 2\sigma$, and $\overline{X}\pm 3\sigma$, we have the percentages 68.268, 95.450, and 99.730 respectively. Obviously, it is highly improbable that the difference between averages should be greater than three times the standard deviation of the average, providing we assume that all of the samples were drawn from the same universe: In other words, that all of the samples were manufactured under the same random conditions. The fourth column, then, indicates practical limits to the variations in the averages. It is obvious, therefore, that the differences between the averages are larger than could have been expected, if the same system of causes controlled the different groups of observations. In other words the differences are significant and must be explained.

Why do these variations exist? We shall show in the course of the discussion that the normal law is not sufficient to answer these questions. We shall show also that the variations noted are largely the result of the method of sampling used at that time. The significance of the other factors given in this table is discussed later.

Why Is the Application of the Normal Law Limited?

Why can we not assume that the deviations follow the normal law of error? This is

$$y=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}} \tag{1}$$

where σ is the root mean square error $\sqrt{\frac{\Sigma yx^2}{n}}$ and y is the frequency of occurrence of the deviation x from the arithmetic mean and n is the number of observations? If they do, the answers to all of the questions raised in the preceding paragraphs can be easily answered in a way which is familiar to all acquainted with the ordinary theory of errors and the method of least squares. This is an old and much debated question in the realm of statistics. Let us review briefly some of the a posteriori and a priori reasons why the normal law has gained such favor and yet why it is one of the most limited, instead of the most general, of the possible laws.

A Posteriori Reasons. The original method of explaining the normal law rests upon the assumption that the arithmetic mean value of the observations is always the most probable. Since experience shows that the observed arithmetic mean seldom satisfies the condition of being the most probable we may justly question the law based upon an apparently unjustified assumption.

Gauss first enunciated this law which is often called by his name. The fact that so great a mathematician proposed it led many to accept it. He assumes that the frequency of occurrence of a given error is a function of the error. The probability that a given set of n observations will occur is the product of the probabilities of the n independent events. He then assumes that the arithmetic mean is the most probable and finds the equation of the normal law. Thus he *assumes* the answer to the first question; that is, he assumes that the most probable value is always the arithmetic mean. In most physical and engineering measurements the deviations from the arithmetic mean are small, and the number of observations is not sufficiently large to determine whether or not they are consistent with the assumption of the normal law. Under these conditions this law is perhaps as good an approximation as any.

The fundamental assumptions underlying the original explanation were later brought into question. What a priori reason is there for assuming that the arithmetic mean is the most probable value? Why not choose some other mean? [6] Thus if we assume that the median [7] value is the most probable, we obtain as a special case the law of error represented by the following equation:

$$y = Ae^{-h^2|x|} \tag{2}$$

where y represents the frequency of occurrence of the deviation x from the median value and e is the Naperian base of logarithms. Both A and h are constants. If, however, we assume that the geometric mean is the most probable, we have as a special case the law of error represented by the following equation:

$$y = Ae^{-h^2(\log X - \log a)^2} \tag{3}$$

where in this case y is the frequency of occurrence of an observation of magnitude X, "a" is the true value, and A and h are constants.[8]

Enough has been said to indicate the significance of the assumption that the arithmetic mean is the most probable value, but, why choose this instead of some other mean? No satisfactory answer is available. So far as the author has been able to discover, no distribution representing physical data has even been found which approaches the median law. Several examples have been found in the study of carbon which conform to the law of error derived upon the assumption that the geometric mean is the most probable. If the arithmetic mean were observed to be the most probable in a majority of cases, we might consider this an a posteriori reason for accepting the normal law. We find the contrary to be the case.

Furthermore, we find in general that the distribution of errors is non-symmetrical about the mean value. In fact, most of the distributions which are given in textbooks dealing with the theory of errors and the method of least squares to illustrate the universality

[6] An average or mean value may be defined as a quantity derived from a given set of observations by a process such that if the observations became all equal, the average will coincide with the observations, and if the observations are not all equal, the average is greater than the least and less than the greatest.

[7] If a series of n observations are arranged in ascending order of magnitude, the median value is that corresponding to the observation occurring midway between the two ends of the series.

[8] A very interesting discussion of the various laws that may be obtained by assuming different mean values is given in J. M. Keynes' "A Treatise on the Theory of Probability."

of the law are, themselves, inconsistent with the assumption of such a law. Prof. Pearson was one of the first to point out this fact. He considers among others an example originally given by Merriman [9] in which the observed distribution is that of 1,000 shots fired at a target. The theoretical normal is the solid line in Fig. 5 and the

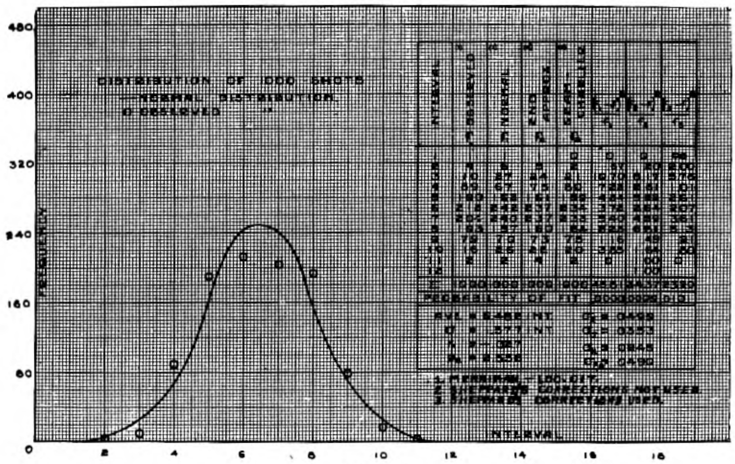


Fig. 5

observed frequencies are the small circles. When represented in this way there appears to be a wide divergence between theory and experience. Of course, some divergence may always be expected as a result of variations due to sampling; and, too, we must always question a judgment based entirely upon visual observation [10] of a graphical representation of this character. Prof. Pearson uses his method—which will be discussed later—for measuring the goodness of fit between the theoretical and observed distributions. He [11] finds that a fit as bad or worse than that observed could have been expected to occur on an average of only 15 to 16 times in ten million. We must conclude, therefore, that these data are not consistent with the assumption of a universal normal law.

A Priori Reasons. From the physicist's viewpoint the origin of the Gaussian law may be explained upon a more satisfactory basis.

[9] "Method of Least Squares," Eighth Edition—Page 14.

[10] This point will be emphasized later:—first, by showing that these data *appear* consistent with a normal law when plotted on probability paper, and second, by showing that some frequency distributions appear normal when plotted even though they are not. The other data in this table will be referred to later.

[11] Reference to the original article and a quotation therefrom given in the eleventh edition of the *Encyclopedia Britannica* on the article "Probability."

It is that which was originally suggested by La Place. If, however, we accept this explanation, we must accept the fact that the normal law is the exception and not the rule. Let us consider why this is true.[12]

This method of explanation rests upon the assumption that the normal law is the first approximation to the frequencies with which different values will be assumed by a variable quantity whose variations are controlled by a large number of independent causes acting in random fashion. Let us assume that:

a. The resultant variation is produced by n causes.
b. The probability p that a single cause will produce an effect Δx is the same for all of the causes.
c. The effect Δx is the same for all of the causes.
d. The causes operate independently one of the other.

Under these assumptions the frequency distribution of deviations of 0, 1, 2 . . . n positive increments can be represented by the successive terms of the point binomial $N(q+p)^n$ where N represents the total number of observations.

Under these conditions if $p=q$ and $n=\infty$, the ordinates of the binominal expansion can be closely approximated by a *normal* curve having the same standard deviation. These restrictions are indeed narrow. In practice it is probable that p is never equal to q, and it is certain that n is never infinite. Therefore, the normal distribution should be the exception and not the rule.

There is a more fundamental reason, however, why we should seldom expect to find an observed distribution which is consistent with the normal law. In what has preceded we have assumed that each cause produced the same effect Δx, and that the total effect in any instance is proportional to the number of successes.

Let us assume that the resultant effect is, in general, a function of the number n of causes producing positive effects, that is, let $X=\phi(n)$. Thus we assume that the frequency distributions of the number of causes and of the occurrence of a magnitude X are respectively

$$y=f(n)$$

and

$$y_1=f_1\,(X)$$

for two values of n, say n and $n+dn$, there will be two values of X, say X and $X+dX$. The number of observations within this interval of n must be the same as that within the corresponding interval of X.

[12] Bowley "Elements of Statistics," Part II.

If the distribution in X is normal such that we have

$$y_1 = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-a)^2}{2\sigma^2}},$$

then

$$y = \frac{1}{\sigma\sqrt{2\pi}} \phi'(n) e^{-\frac{[\phi(n)--a]^2}{2\sigma^2}} \tag{4}$$

where a is the arithmetic mean value, therefore, the distribution of the causes need not be normal; conversely if the causes are distributed normally, the observations will not in general be normal.

This idea is of great importance in the interpretation of observed distributions of physical data.[13] To illustrate, let us assume that the natural causes which affect the growth of apples on a given tree produce a normal variation in the diameters of the apples. Obviously, the distribution of either the cross-sectional areas or the volumes will not be normal.[14]. If the distribution of the diameters is normal as supposed, the arithmetic means of these diameters is the most probable value. Obviously, however, neither the arithmetic mean area nor the arithmetic mean volume will be the most probable, because in general

$$\frac{1}{n}\Sigma f(X) \neq f(\frac{1}{n}\Sigma X). \tag{5}$$

As already indicated, the deviations dealt with in the present investigation were not small. The form of the observed distribution may be expected, therefore, to depend upon the functional relationship between the observed quantity and the number of causes. We shall

[13] Kapteyn, J. C.—Skew Frequency Curves—Groningen, 1903.

[14] In the theory of errors this fact is taken into account by assuming that the variations are always *small*. Thus, if the variable X can be represented as a function F of certain other variables $U_1, U_2, \ldots U_m$ so that we have

$$X = F(U_1, U_2, \ldots U_m),$$

we ordinarily assume that we can write this expression in the following form

$$X = F(a_1 + u_1, a_2 + u_2, \ldots a_m + u_m).$$

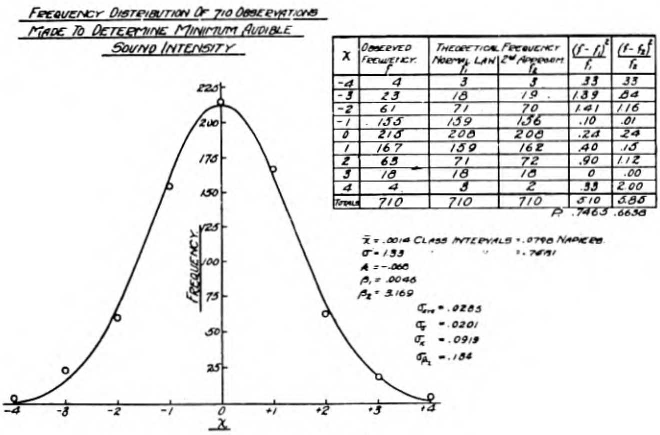
A further assumption is made that the u's are small so that 2nd and higher powers and products of these can be neglected. Under these conditions the distribution of X is normal and has a standard deviation given by the following expression:

$$\sigma_X = \sqrt{\left(\sigma_{U_1}\frac{\partial F}{\partial U_1}\right)^2 + \left(\sigma_{U_2}\frac{\partial F}{\partial U_2}\right)^2 + \ldots \left(\sigma_{U_m}\frac{\partial F}{\partial U_m}\right)^2}.$$

But, thus, we are led to overlook the significance of the form of F, particularly in those practical cases such as are of interest in the present paper where the quantities $u_1, u_2, \ldots u_m$ are not small.

illustrate the significance of these ideas as an aid in the interpretation of data by reference to the results of our study of the law of error of the human ear in measuring the efficiency of transmitters.

Let us consider the problem of determining the minimum audible sound intensity. Let us assume that there are n physiological and psychological causes controlling this sensation measure, and that the probabilities of the causes producing 0, 1, 2 . . . n effects are dis-



X	Observed Frequency f	Theoretical Frequency Normal Law f_1	Theoretical Frequency 2nd Approxim. f_2	$\frac{(f-f_1)^2}{f_1}$	$\frac{(f-f_2)^2}{f_2}$
-4	4	3	3	.33	33
-3	23	18	19	1.39	.84
-2	61	71	70	1.41	1.16
-1	155	159	156	.10	.01
0	215	208	208	.24	24
1	167	159	162	.40	.15
2	63	71	72	.90	1.12
3	18	18	18	0	.00
4	4	3	2	.33	2.00
Totals	710	710	710	5.10	5.85
			P.	.7463	.6638

Fig. 6

tributed normally. Because of these differences in human ears different amounts of sound energy are required to produce minimum audible sensations. What is the distribution of energies?

The data are given in Fig. 6. These have been previously reported by Fletcher and Wegel of this laboratory.[15] The method of making these measurements was described in their original papers. It is sufficient to recall that the results are given in terms of pressures in dynes per square centimeter. Seven hundred and ten observations covering the frequency range of from 60 to 8,000 cycles are included. The data include results for both ears of 14 women and 20 men, and one ear only for two women and two men. Only ears that had been medically inspected as being physiologically normal were selected. These results, therefore, include variations in the observations of a single observer with those of different observers.

The natural logarithms of the intensities were added and the average of these was obtained. The distribution of the natural

[15] Fletcher, H. and Wegel, R. L.—Proceedings of the National Academy of Science—Vol. VIII, pp. 5—6, January, 1922.
Physical Review, Vol. XIX, pp. 550 seq. 1922.

logarithms of the intensities is given in the second column of the table in Fig. 6. The smooth line is the normal curve based upon the observed value of standard deviation. The distribution of the logarithms of the intensities is normal.[16] The arithmetic mean of the logarithms is the most probable. Therefore, the distribution of intensities is decidedly skew, and the geometric mean intensity is the most probable. Here, then, is an excellent example in which it is highly probable that the distribution of the causes is random and normal, but in which the resultant effect is not a linear function of the number of causes.[17]

Can We Ever Expect to Find a Normal Distribution in Nature?

The answer is affirmative. If the resultant effect of the independent causes is proportional to their number, the distribution rapidly approaches normality as the number of causes is increased even though $p \neq q$.

To show this, let us assume that the variation in a physical quantity is produced by 100 causes, and that each cause produces the same effect Δx. Also, let us assume the probability p to be 0.1, that each cause produces a positive effect. The distribution of 0, 1, 2, . . . n successes in 1000 trials is given by the terms of the expansion 1000 $(.9+.1)^{100}$. Obviously such a distribution is skew, p is certainly not equal to q, and n is far from being infinite. If the normal law

[16] In fact this is an exceptionally close approximation to the normal law. This will be more evident after we have considered the methods for measuring the goodness of fit as indicated by the other calculations given in this figure. For the present it is sufficient to know that approximately 75 times out of 100 we must expect to get a system of observations which differ as much or more from the theoretical distribution calculated from the normal law than the observed distribution differs therefrom in this case. The fact that the second approximation does not fit the observed distribution as well as the normal—*i.e.* the measure of probability of fit P is less—indicates that the observed value of the skewness k is not significant.

[17] These results are of particular interest to telephone engineers. The fact that the distribution of the logarithms of the intensities is normal is consistent with the assumption of Fechner's law which states that the sensation is proportional to the logarithm of the stimulus. The range of variation (that is, $\bar{X} \pm 3\sigma$) in different observers' estimates of the sound intensity required to produce the minimum audible sensation is approximately 20 miles. The range of error of estimate depends upon the intensity of sound and decreases as the sound energy level increases. Thus for the average level which prevails for transmission over the present form of telephone system in a three mile loop common battery circuit it is less than 9 miles. Even at this intensity, however, it is obvious that although scarcely any observers will differ in their estimates by more than 9 miles, 50% of them will differ by at least 2 miles. These results also furnish experimental basis for the statement made in the beginning of this paper: that is, the variations introduced in the method of measurement of transmitter efficiencies are large in comparison with the average efficiency.

TABLE III—COMPARSION OF THE TERMS OF THE EXPANSION $1000(.1+.9)^{100}$ WITH THOSE OBTAINED BY VARIOUS APPROXIMATIONS

Number of Successes	* $1000(.1+.9)^{100}$ f	Normal Law f_1	2nd Approximation f_2	*Law of Small Numbers f_3	*Gram Charlier f_4	*Poisson Charlier f_5	† Pearson f_6	$\frac{(f-f_1)^2}{f_1}$	$\frac{(f-f_2)^2}{f_2}$	$\frac{(f-f_3)^2}{f_3}$	$\frac{(f-f_4)^2}{f_4}$	$\frac{(f-f_5)^2}{f_5}$	$\frac{(f-f_6)^2}{f_6}$
−1	.0	.2	.0	.0	.0	.0	.0	.200	.0	.0	.0	.0	.0
0	.1	.6	.0	.0	.0	.0	.9	.417	∞	∞	∞	∞	.711
1	.3	1.5	.4	.5	.4	.2	2.2	.960	.025	.080	.025	.050	1.641
2	1.6	3.9	2.1	2.3	2.0	1.6	7.1	1.356	.119	.213	.080	.0	4 261
3	5.9	8.9	6.7	7.6	6.5	5.9	15.9	1.011	.096	.380	.055	.0	6.289
4	15.9	18.3	16.7	18.9	16.2	15.9	28.9	.315	.038	.476	.006	.0	5.848
5	33.9	33.4	34.0	37.8	33.3	34.0	44.8	.007	.0	.402	.011	.0	2.652
6	59.6	54.8	58.8	63.0	58.1	59.9	61.3	.420	.011	.183	.039	.002	.047
7	88.9	80.9	88.0	90.1	87.5	89.3	76.1	.791	.009	.016	.022	.002	2.153
8	114.8	106.0	113.9	112.5	114.5	114.9	87.0	.731	.007	.047	.001	.0	8.883
9	130.4	125.2	130.5	125.1	131.8	130.1	92.8	.216	.0	.225	.015	.001	15.234
10	131.9	132.6	132.6	125.1	133.9	131.4	93.3	.004	.004	.370	.030	.002	15.970
11	119.9	125.2	119.9	113.7	121.1	119.4	89.1	.224	.0	.338	.012	.002	10.647
12	98.8	106.0	98.1	94.8	98.4	98.6	81.2	.489	.005	.169	.002	.0	3.815
13	74.3	80.9	73.8	72.9	73.2	74.4	71.0	.538	.003	.027	.017	.0	.153
14	51.3	54.8	50.8	52.1	50.2	51.4	59.9	.224	.005	.012	.024	.0	1.235
15	32.7	33.4	32.8	34.7	32.2	33.0	48.7	.015	.0	.115	.008	.003	5.257
16	19.3	18.3	19.9	21.7	19.4	19.7	34.5	.055	.018	.265	.001	.008	6.697
17	10.6	8.9	11.1	12.8	10.9	10.8	29.5	.325	.023	.378	.008	.004	12.109
18	5.4	3.9	5.7	7.1	5.8	5.5	22.0	.577	.016	.407	.028	.002	12.525
19	2.6	1.5	2.6	3.7	2.7	2.5	16.0	.807	.0	.327	.004	.004	11.223
20	1.2	.6	1.3	1.9	1.2	1.1	11.4	.600	.008	.258	.0	.009	9.126
21	.5	.1	.4	.9	.5	.5	7.9	1.600	.025	.178	.0	.0	6.932
22	.2	.1	.2	.4	.2	.1	5.4	.100	.0	.100	.0	.1	5.007
23	.0	.0	.0	.2	.0	.0	3.6			.200			3.600
24	.0	.0	.0	.0	.0	.0	2.3						2.300
25	.0	.0	.0	.0	.0	.0	1.5						1.500
26	.0	.0	.0	.0	.0	.0	.9						.900
27	.0	.0	.0	.0	.0	.0	.6						.600
28	.0	.0	.0	.0	.0	.0	.4						.400
29	.0	.0	.0	.0	.0	.0	.2						.200
30	.0	.0	.0	.0	.0	.0	.1						.100
Σ	1000.1	1000.0	1000.3	999.8	1000.0	1000.2	996.5	11.982	.412	5.166	.388	.189	158.015
Ave.	9.998	10.000	9.996	9.999	9.998	10.000	10.754						

$p=.1$ $\quad \sigma=\sqrt{pqn}=3$ $\quad \beta_2=3+\frac{1-6\,pq}{pqn}=3.511$ $\quad \sigma_{\beta_2}=\sqrt{\frac{24}{1000}}=.155$

$q=.9$

$n=100$ $\quad K=\frac{q-p}{\sqrt{pqn}}=.267$ $\quad \sigma_k=\sqrt{\frac{6}{1000}}=.077$

$pn=10$

* Fisher, A.—"The Mathematical Theory of Probabilities."

† Pearson, Karl—Phil. Mag., 1907, pp. 365–378.

were fitted to such a distribution, would it be possible to detect easily any great difference between theory and observation?

Let us compare the two distributions. The data are given in Table III. First, the average value must be the most probable in order to be consistent with the normal law. It is, because the observed most probable value corresponds to 10 successes, and the average of

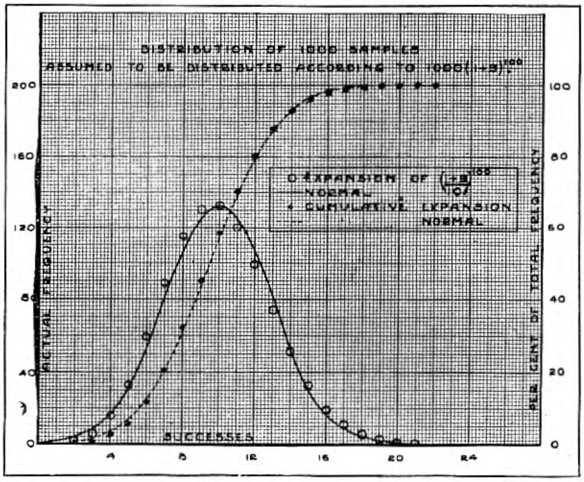


Fig. 7

the hypothetically observed distribution is 9.998. This under ordinary circumstances would be considered a close check between theory and practice.

The normal distribution is given in the third column of the table. Even though there is a difference between the frequencies given in the second and third columns, would the average observer be apt to conclude that the hypothetically observed distribution is other than normal? He would probably base his answer upon a graphical comparison such as given in Fig. 7. The solid line represents the normal curve; whereas the frequencies given in the second column of Table III are represented by circles. It is obvious that the normal law appears to be a very close approximation to the terms of the binomial expansion.

Thus we see that for even a small number of causes the difference between p and q may be quite large, and yet the difference between the distributions given by the binomial expansion and that given by the normal law is apparently small and not easily to be detected by ordinary methods. As n increases the closeness of fit does likewise.

If p is equal to q, the number of causes must be very small indeed before we are able to detect the difference between the terms of the binomial expansion and those given by the normal law. To show that this is true I have chosen a case corresponding to a physical condition where there are only 16 causes and where p is equal to q. The data are given in Table IV.

TABLE IV

Successes	$(.5+.5)^{16}$ f	Normal Law with same σ f_1
0	.0000153	.0000669
1	.0002441	.0004363
2	.0018311	.0022159
3	.0085449	.0087641
4	.0277710	.0269955
5	.0666504	.0647586
6	.1220825	.1209853
7	.1745605	.1760326
8	.1963806	.1994711
9	.1745605	.1760326
10	.1220825	.1209853
11	.0666504	.0647588
12	.0277710	.0269955
13	.0085449	.0087641
14	.0018311	.0022159
15	.0002441	.0004363
16	.0000153	.0000669

Obviously, therefore, the limitations imposed by the assumptions as to the number of causes and the equality of p and q are not as important as they might at first appear. It is probable that this is one of the reasons why we find approximately normal distributions. If, however, p is sufficiently small, the difference between the observed distribution and that consistent with the normal law can easily be detected. We shall show in a later section that this is true for Rutherford's data.[18]

Is There a Universal Law of Error ?

Obviously from what has already been said, the normal law is not a universal law of nature. It is probable that no such law exists. We do, however, have certain laws which are more general than the normal. We shall consider briefly some of these types in an effort to indicate the advantages that can be gained by an application of them to physical data.

[18] Loc. cit.

Binomial Expansion $(p+q)^n$. We have already seen that the distribution is approximately normal when $p=q$ and $n \doteq \infty$. Following Edgeworth[19], Bowley [20] shows that if $p \neq q$ but $n \doteq \infty$ the frequency y of the occurrence of a deviation of magnitude x is given by the following expression where k represents the skewness [21] of the distribution:

$$y = \frac{1}{\sigma\sqrt{2\pi}}\left(\exp. -\frac{x^2}{2\sigma^2}\right)\left[1 - \frac{k}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right]. \tag{6}$$

This will be referred to as the *second approximation.*

If p is very small, but $pn = \lambda$ is finite, we have the so-called *law of small numbers* [22] which was first derived by Poisson. The successive terms of the series $e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{3}+\ldots\right)$ represent the chances of 0, 1, 2 . . . n successes. Theoretically, if we are dealing with a distribution of attributes,[23] it is always possible to calculate the values of

[19] Cambridge Philosophical Transactions, Vol. XX, 1904, pp. 36-65 and 113-141.

[20] Loc. cit.

[21] In statistical work the practice is followed of using the moments of the distribution for determining the parameters of the frequency curve. The i th moment μ_i of a frequency distribution about the arithmetic mean is by definition

$$\mu_i = \frac{\Sigma y x^i}{\Sigma y}$$

In calculating such moments it is necessary to consider the observations as grouped about the mid-point of the class interval and unless this interval is very small certain errors are introduced which can be partially eliminated by applying Sheppard's corrections as given by him in Biometrika, Vol. III, pages 308 seq. If Δx be taken as unity, we have

$$\mu_1 = 0 \qquad \beta_1{}^2 = k = \frac{q-p}{\sqrt{pqn}}$$
$$\mu_2 = pqn = \sigma^2$$
$$\mu_3 = pqn(q-p)$$
$$\mu_4 = 3(pqn)^2 + pnq(1-6pq) \qquad \beta_2 = 3 + \frac{1-6pq}{pqn}$$

and if p is approximately equal to q and n is large we have $\sigma_k = \sqrt{\frac{6}{N}}$ and $\sigma_{\beta_2} = \sqrt{\frac{24}{N}}$.

[22] It is of interest to note that several investigators have derived this law independently. Thus H. Bateman derives this expression in an appendix to the article of Prof. Rutherford and H. Geiger previously referred to. This is, in a way, an illustration of the apparent need of a broader dissemination of information relating to the application of statistical methods of analysis to engineering and physical data. It is also of interest to note that this law has been used to advantage in the discussion of telephone trunking problems.

[23] If the classification is based upon the presence or absence of a single characteristic, this characteristic is often referred to as an attribute.

p, q and n from the moments of the distribution.[24] Even when p, q and n are known, the arithmetic involved in calculating the terms of the binomial is often prohibitive, and, therefore, it is necessary to obtain certain approximations corresponding to the three laws of error; that is, normal, second approximation, and the law of small numbers. Tables for the normal law and for the law of small numbers are readily available in many places, while those for the second approximation are given by Bowley.[25]

Even under conditions where the binomial expansion does not hold, Edgeworth has shown that it is possible to obtain the following general approximation:

$$y=\frac{1}{\sigma\sqrt{2\pi}}\left(\exp.-\frac{x^2}{2\sigma^2}\right)\left[1-\frac{k}{2}\left(\frac{x}{\sigma}-\frac{x^3}{3\sigma^3}\right)+\frac{k^2}{8}\left(-\frac{5}{3}+\frac{5x^2}{\sigma^2}-\frac{5x^4}{3\sigma^2}+\frac{x^6}{9\sigma^6}\right)\right.$$
$$\left.+\left(\frac{\mu_4-3\mu^2}{8\sigma^4}\right)\left(1-\frac{2x^2}{\sigma^2}+\frac{x^4}{3\sigma^4}\right)\right]. \quad (7)$$

This holds providing the observations are influenced by a large number of causes, each of which varies according to some law of error but not necessarily to the normal law.

Gram-Charlier Series. Gram, according to Fisher,[26] was the first to show that the normal law is a special case of a more generalized system of skew frequency curves. He showed that the arbitrary frequency function $F(X)$ can be represented by a series of terms in which the normal law is the generating function $\phi\ (X)$. Thus

$$F(X)=c_0\phi(X)+c_1\phi'(X)+c_2\phi''(X)+\ldots \quad (8)$$

where c_0, c_1, c_2, etc., are constants which may be determined from the moments of the observed data. This series is similar to that already mentioned in the above equation (7) which Edgeworth has obtained in several different ways. This law is of interest from the viewpoint of either a physicist or an engineer in so far as it gives him a picture of the casual conditions consistent with an accepted theoretical curve. Thus, if either the causes of variation are within a certain degree not entirely independent, or the errors are not linearly aggregated, the observed frequency distributions may be expected to conform to an equation such as 8. This equation has been found to fit a much larger group of observed distributions than the normal law

[24] See footnote 26.

[25] See for example Pearson, K.—Tables for Biometricians and Statisticians—Cambridge University Press.

[26] Fisher, Arne—Theory of Probabilities—page 182.

and the publication of the necessary tables by Fisher [27] and Glover [28] makes the study of such a curve more feasible. The author finds, for example, that this series furnishes a much closer fit to the distribution of shots, Fig. 5, referred to above than any other that he has tried.

Theoretically we should be able to improve the approximation by taking a large number of terms of the series. Such a procedure, however, involves the use of moments higher than the first four, and the errors in these moments are so large as to make their use impractical.

In spite of the uncertainty attached to the interpretation of the physical significance of fitting any of these curves to data, one very practical observation has been made: that is, if an observed series of frequencies could not be fitted by a theoretical curve in any of the ways already mentioned, careful consideration of the possible reasons for the observed poor fit have in practically every instance suggested the cause or causes thereof. We shall refer to only one practical example.

The data have already been given above in Table II. It has been noted that in this instance the variations in the averages of groups of several thousand observations showed that the differences were significant. If the observed distributions had been normal, it would have been necessary to assume either that the methods of making the measurements were different for the different groups of observers, and for the different machines, or that the manufacturing methods were experiencing a trend. Although the observed frequency curves for the different groups were found to be smooth, the observed frequencies could not be readily fitted by any curve previously described. This naturally led to a search for the existence of any one of a number of causes affecting the observations which might produce such a divergence between theory and practice. One by one these causes were found and eliminated and as they were the degree of fit between the results of theory and practice increased. For example, it was found that some of the groups of observations were for transmitters assembled from only two or three lots of carbon. Transmitters assembled from one lot of carbon had a different average efficiency from those assembled from another lot. Naturally the

[27] Fisher, Arne—Loc. cit. As noted by Mr. Fisher, page 214, the values of $\phi(x)$ and its first 6 derivatives to 7 decimal places for values of x up to 4 and progressing by intervals of 0.01 were given by Jörgensen in his "Frekvensflader og Korrelation."

[28] Glover, J. W.—Tables of Compound Interest, Functions, etc.—1923 Edition published by George Wahr, Ann Arbor, Michigan.

resultant distribution was a compound of a few separate but similar distributions about different averages. When the distributions of the efficiencies of the different lots of carbon were determined separately they were found to be consistent with the second approximation.

Thus, although it may be impossible to conclude that the a priori assumptions underlying a given law of distribution are fulfilled because the observations are found to be consistent therewith, nevertheless, the fact that the observed and the theoretical distributions do not agree suggests the necessity of seeking for certain typical causes which may be expected to introduce such discrepancies. This point is of special importance in connection with the study of ways of sampling product in order to determine whether or not the manufacturing process is subject to trends. Thus, if a product is sampled at two periods, and the distributions of both groups of observations are found to be random about different averages, it is highly probable that the difference indicates a trend in the manufacturing methods, providing the difference between the averages is greater than 3 times the standard deviation of the average. When, however, the two distributions are found to be inconsistent with a random system of causes, it is quite probable that the condition of sampling has not been carefully controlled.

Hypergeometric Series. Pearson has shown several ways in which a frequency distribution may be represented by a hypergeometric series. Thus the chances of getting r, $r-1, \ldots 0$ bad transmitters from a lot containing pn bad and qn good and where r instruments are drawn at a time may be represented by the terms of such a series. More important, however, is Pearson's solution [29] of what he calls the fundamental problem of statistics. He shows, following the line of reasoning similar to that originally suggested by Bayes, that if in a sample of $k_1=(m+n)$ trials, an event has been observed to occur m times and to fail n times, in a second group of k_2 trials the chances of the event occurring r times and failing s times are given by the successive terms of a hypergeometric series. We cannot consider here the questions underlying the justification of this method of solution, for, as is well-known, the application of Bayes' theorem is questioned by many statisticians. We can profit, however, by the broad experience of Prof. Pearson, for he has apparently accumulated an abundance of data which are consistent with the theory.

The answer to this problem s of special importance in connection with the inspection of product which in many instances runs into millions yearly. We must keep the cost of inspection at a minimum,

[29] Pearson, K.—Biometrika, October, 1920—pp. 1-16.

which means that the sample numbers must be small, and yet we see from the solution derived from Pearson the significance of the sizes of both the original and the second sample. Thus, he [30] shows that the standard deviation σ is given by the equation

$$\sigma^2 = k_2 pq\left(1 + \frac{k_2}{k_1}\right). \tag{9}$$

Multimodal Distributions. These occur frequently in engineering work and particularly in connection with the inspection of large quantities of apparatus. One such instance has already been referred to in the discussion of the data given in Table II, and another is illustrated by the data given in Fig. 1. Prof. Pearson [31] has developed a method for determining analytically whether or not the observed distribution is such as may be expected to have arisen from the combination of two normal components, the mean values of which are different. The method involves the solution of a ninth degree equation. As a result, the arithmetic work is in many cases prohibitive. This method cannot be applied to the data given in Fig. 1 primarily because the number of observations is not sufficiently great.

Pearson's Closed Type Curves.[32] One of the best known statistical methods for graduating data is that developed by Prof. Pearson. His system of closed type curves arises from the solution of the differential equation derived upon the assumption that the distribution is uni-modal and touches the axis when $y=0$. In the hands of Pearson and his school great success has been attained in graduating data collected from widely different fields, although primarily from these of biology, psychology, and economics. The choice of curve to represent a given distribution rests primarily upon a consideration of a criterion involving two constants, $\beta_1 = \sqrt{k}$ and β_2, both of which have been defined previously in footnote 21.

In the early study of the distributions of efficiencies of product transmitters an attempt was made to apply this system of curves. For example, the Pearson types are indicated in Table II. In no instance, however, was it possible to obtain a very satisfactory fit between the observed and the theoretical distributions. Furthermore, the arithmetical work required to calculate a theoretical distribution in this way is excessive. We must also consider what physical significance can be attached to the different types of curves. The answer is not definite. Under certain conditions the generalized

[30] Pearson, K.—*Philosophical Magazine*—1907, pp. 365-378.

[31] Pearson, K.—*Philosophical Magazine*—Vol. 1, 1901, pp. 115-119.

[32] Elderton—Frequency Curves and Correlation.

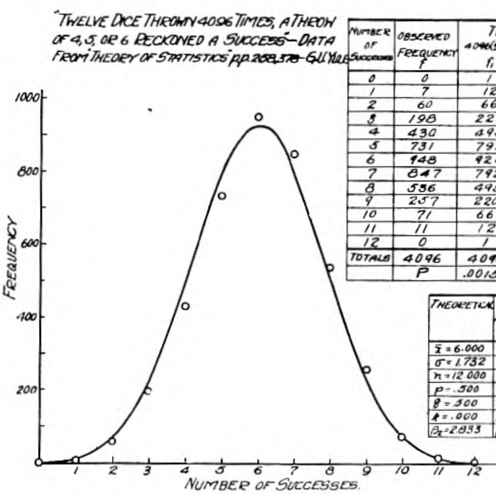
equation of Pearson breaks down to the normal law and the second approximation. These, of course, can be explained as previously. The fundamental equation, however, serves to cover the condition where the causes are correlated. Thus, because of the lack of a clear conception of the physical significance of the observed variations in the type of curves indicated in Table II, it was not possible easily to set up experiments to find the causes of these variations. For this reason preference has been given to the use of frequency distributions derived upon a less empirical basis following the original lines laid down by La Place, Edgeworth, Kapteyn, and others previously referred to. Another very practical reason for choosing the latter type of curve is that it involves for the most part the use of only the first three moments of the distribution instead of the first four required for differentiating between the Pearson types. In those cases where the interest is less of physical interpretation than of graduating an observed set of data, preference may go to the more generalized system of Pearson.

How Can We Choose the Best Theoretical Frequency Distribution?

We have already briefly reviewed some of the different methods for obtaining a theoretical frequency distribution from a consideration of the moments of the observed frequencies. We have seen in Table III that by using different methods we obtain different degrees of approximation to the hypothetically observed distribution which in this case corresponds to the terms of the binomial expansion $1000(.1+.9)^{100}$. Similarly from Fig. 5 it is seen that the Gram-Charlier series is a much closer approximation to the observed distribution than that derived upon the assumption of the normal law. In any given case we are naturally confronted with the question: What is the best theoretical distribution? We shall consider four methods for obtaining an answer.

The oldest, simplest, and in many instances the most practical, is that of comparing graphically or in tabular form the theoretical distribution with the one observed. This method is, however, inaccurate and qualitative. It does not furnish us with a quantitative method of measuring the closeness of fit between theory and practice, and in certain instances it is absolutely misleading. It is of interest to see how all of these things can be truly said of one and the same method. The first two characteristics, that is, oldest and simplest, are perhaps readily granted. It remains to be pointed out more definitely wherein the method is sadly deficient as a quantitative measure, and therefore often misleading; whereas in certain instances it may be, nevertheless, the only practical method that can be used.

Graphical Method. The graphical method itself may be subdivided into two parts. Let us consider first the plot of the observed and theoretical frequencies. As an example of the unsatisfactory nature of this form of comparison, it is of interest to consider certain data



NUMBER OF SUCCESSES	OBSERVED FREQUENCY f	THEORETICAL FREQUENCY $4096(\frac{1}{2}+\frac{1}{2})^{12}$ f_1	NORMAL LAW f_2	2ND APPROXIM. $k=-.061$ f_3	$k=-.027$ f_4
0	0	1	2	2	2
1	7	12	12	14	13
2	60	66	55	58	57
3	198	220	184	185	184
4	430	495	441	453	437
5	731	792	759	747	754
6	948	924	958	936	957
7	847	792	832	843	837
8	536	495	530	540	535
9	257	220	242	243	242
10	71	66	80	77	79
11	11	12	18	16	17
12	0	1	3	2	3
TOTALS	4096	4096	4096	4096	4097
	P	.0015	.2409	.4557	.3062

THEORETICAL	CALCULATED FR 1ST + 2ND MOMENTS.	FR 1ST, 2ND, 3RD + 4TH MOMENTS.
$\bar{x}=6.000$	$\bar{x}=6.139$	$\bar{x}=6.139$
$\sigma=1.732$	$\sigma=1.712$	$\sigma=1.712$
$n=12.000$	$n=11.738$	
$p=.500$	$p=.523$	
$q=.500$	$q=.477$	
$k=.000$	$k=-.027$	$k=-.061$ $PE_k=.026$
$\beta_2=2.833$	$\beta_2=2.850$	$\beta_2=2.787$ $PE_{\beta_2}=.052$

Fig. 8

given by Yule [33] in which 12 dice are thrown 4,096 times, a throw of 4, 5, or 6 points being reckoned a success. If the dice are symmetrical $p=q=\frac{1}{2}$ and the theoretical distribution if given by 4,096 $(\frac{1}{2}+\frac{1}{2})^{12}$, the terms of which as given by Yule are presented in the third column of Fig. 8. It is suggested that the reader, before going further, consider the graphical and tabular representation of these data. The smooth curve is the theoretical distribution $4{,}096(\frac{1}{2}+\frac{1}{2})^{12}$. It has been the author's experience to find that in practically every instance in which this curve has been shown to an individual for the first time that the impression is that which Yule evidently desires to produce by the illustration: that is there is a very good fit between theory and practice. This distribution is, however, not symmetrical: it is skew. The dice used in this experiment were not symmetrical: that is, $p \neq q$. How do we know that these statements are true?

Let us consider the normal and second approximation as given

[33] Yule—"Introduction to the Theory of Statistics."

in the fourth, fifth, and sixth columns.[34] Obviously the degree of fit is closest for the second approximation, although that between the normal distribution and the observed frequencies is closer than that between the terms of the binomial expansion and the observed frequencies. To be sure, the normal law is only an approximation to the point binomial when $p=q$ and $n=\infty$. The normal distribution, however, is calculated about the observed average 6.139, instead of about the theoretical average 6. If the dice are non-symmetrical, the average will not be 6, and, therefore, the center of the distribution will be shifted after the fashion observed. The improvement in fit corresponding to the normal distribution is therefore primarily attributable to that introduced by shifting the center of the distribution indicating that $p \neq q$. However, if $p \neq q$, the second approximation should improve the fit and for either value of k this is found to be the case. Thus even though we cannot measure quantitatively the improvement of fit, the qualitative evidence presented in this figure is sufficient to warrant the conclusion that the dice were non-symmetrical, and therefore, that the smooth curve is an unsatisfactory graduation of the data. In fact, by using a quantitative method for measuring the goodness of fit to be discussed in a succeeding paragraph, it follows that only 15 times out uf 10,000 can we expect a divergence from theory as large or larger than that exhibited by the frequencies corresponding to the point binomial.

We have also previously called attention to the fact that in Fig. 7 the eye does not serve to differentiate satisfactorily between the distribution calculated upon the assumption of the normal law and that given by the binominal expansion when the conditions underlying the normal law are far from being satisfied.

Regardless of these criticisms, such graphical methods cannot be entirely dispensed with. Thus the graphical representation of the data given in Fig. 1 shows very clearly that the distribution is probably bimodal, although with no more observations than are available it is practically impossible to show that this is true in any other way.

Instead of plotting the frequency y of occurrence of a variable of magnitude x as ordinate, and x as abscissa, the practice is often followed of plotting as ordinate the percentage of the total number N of observations having magnitudes of x or less.[35]

Any curve $\phi\ (y, x)=0$ may be replaced by a straight line.[36] In

[34] Two values of k were calculated as indicated in the lower right hand corner of the figure.

[35] Heindlhoffer, K. and Sjövall, H.—Endurance Test Data and their Interpretation—Advance paper presented at the Meeting of the American Society of Mechanical engineers, Montreal, Canada, May 28 to 31, 1923.

[36] Runge, C.—*Graphical Methods,* p. 53.

this way we can transform the integral curve into a straight line by choosing an x-scale proportional to the integral from 0 to x of the probability curve.[37] When plotted in this way, a normal distribution appears as a straight line on such paper. At first it may appear very simple to determine whether or not the data conform to a straight line, but in practice this is not always so easy. Thus, we have seen that the distribution of shots presented in Fig. 5 is not normal, but

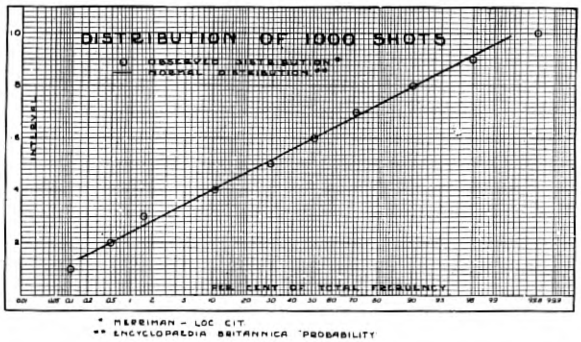


Fig. 9

when these results are plotted on probability paper we have the curve given in Fig. 9. The reader should be cautioned that in such a case there is a temptation to consider that the observed points are approximately well fitted by the straight line, although this is not the case.

Probability paper could be ruled for different theoretical distributions, but in its present form it serves only to determine whether or not the distribution is approximately normal. Its use leaves much to be desired in the way of a quantitative measure of the degree of fit between the theoretical and observed distributions.

Calculation of σ, $\beta_1 = \sqrt{k}$, *and* β_2. Let us consider what information can be obtained as to the best theoretical distribution from only a consideration of the first four moments of the observed frequencies. Let us consider the values of k and β_2 presented in Table V. These have been calculated for the point binomial $(p+q)^n$ where p, q and n have been given different values. For the normal law corresponding to $p=q$ and $n=\infty$, we have $k=0$ and $\beta_2=3$. Thus, if in a practical

[37] Whipple, G. C.—The Elements of Chance in Sanitation—*Franklin Institute Journal,* Vol. 182, July, December, 1916—pp. 37-59 and 205-227.

TABLE V

p	$n=4$		$n=9$		$n=16$		$n=25$		$n=100$		$n=10,000$	
	k	β_2	k	β_2	k	β_2	k	β_2	k	β_2	k	β_2
.5	0	2.50	0	2.78	0	2.87	0	2.92	0	2.98	0	3.00
.6	−.20	2.54	−.14	2.80	−.10	2.89	−.08	2.93	−.04	2.98	−.004	3.00
.7	−.44	2.69	−.29	2.86	−.22	2.92	−.17	2.95	−.09	2.99	−.009	3.00
.8	−.75	3.06	−.50	3.03	−.38	3.02	−.30	3.01	−.15	3.00	−.015	3.00
.9	−1.33	4.28	−.69	3.57	−.67	3.32	−.53	3.20	−.27	3.05	−.027	3.00
.99	−4.92	26.75	−3.28	13.56	−2.46	8.94	−1.97	6.80	−.98	3.95	−.098	3.01
.999	−15.79	251.75	−10.52	113.45	−7.90	65.19	−6.31	42.76	−3.16	12.95	−.316	3.10
.9999	−50.00	2501.75	−33.33	1113.44	−25.00	627.69	−20.00	402.76	−10.00	102.94	−1.000	4.00

case we find an observed distribution for which $k=0$ and $\beta_2=3$, it is highly probable that the distribution is approximately normal. It is true, however, that in sampling from a universe in which $p=q$ and $n=\infty$, the observed values of k and β_2 will seldom be exactly equal to 0 and 3 respectively. Then we must ask what range of values may be expected in these two factors for distributions which are practically normal. For such cases the variations in k and β_2 are practically normal [38] and have standard deviations $\sigma_k=\sqrt{\frac{6}{N}}$ and $\sigma_{\beta_2}=\sqrt{\frac{24}{N}}$ where N is the number of observations. Thus, theoretically any series of observations for which the calculated values of k and β_2 fall within the ranges $0\pm3\sigma_k$ and $3\pm3\sigma_{\beta_2}$ may have arisen from a normal universe. Since, however, the errors σ_k and σ_{β_2} of sampling are so large, this method does not furnish a very practical test for distribution consisting of only a few observations. This is particularly true since, even for very skew distributions, the values of k and β_2 do not differ much from 0 and 3 respectively (see Table V). If, however, the number of observations is large, the values of k and β_2 in themselves often indicate very definitely that the observed frequencies are not consistent with the normal law. For example the calculated values of k and β_2 given for the inspection data in Table II show conclusively that in practically every instance the observed data could not have arisen from a normal universe. So long as we do not use Pearson's system of curves, all that these two factors indicate is that the observed data do or do not conform to the normal law and in this respect their use is limited as is that of the probability paper mentioned above.

In order to show that the factor β_2 is not in itself a very sensitive measure of the variability from the normal law, I have considered the following special case. Let us assume that the observed distributions can be grouped into two parts depending upon whether or not the observations cluster about the average $\overline{X}_1$ or $\overline{X}_2$ measured from a point which is the arithmetic mean of the entire distribution taken about a common origin. This corresponds to the practical case such as that indicated by Fig. 1 which as already pointed out often occurs in practice.

[38] For a critical study of the conditions under which the probable errors of these constants have a real significance, reference should be made to a discussion of this problem by Isserlis in the Proceedings of the Royal Society, series A, Vol. 92, pp. 23 seq.—1915. Obviously even for the normal distribution all of the moments will be skew. This follows from a consideration of equation 4.

The value of β_2 for the entire distribution is then given by the following expression:

$$\beta_2=\frac{(\bar{X}_1{}^4\sum y_1+\bar{X}_2{}^4\sum y_2)+6(\bar{X}_1{}^2\sigma_1{}^2\sum y_1+\bar{X}_2{}^2\sigma_2{}^2\sum y_2)}{(\sum y_1+\sum y_2)\mu_2{}^2}$$

$$+\frac{4(\bar{X}_1\ {}_1\mu_3\sum y_1+\bar{X}_2\ {}_2\mu_3\sum y_2)+({}_1\mu_4\sum y_1\ {}_2\mu_4\sum y_2)}{(\sum y_1+\sum y_2)\mu_2{}^2},$$

where ${}_1\mu_i$ and ${}_2\mu_i$ refer to the adjusted ith moments of the observations about their respective mean values. Let us assume that $\bar{X}=\bar{X}_1=\bar{X}_2$; $k_1=k_2=0$; ${}_1\beta_2={}_2\beta_2=3$; ${}_1\mu_i={}_2\mu_i$; $\sum y_1=\sum y_2$; and $\sigma_1=\sigma_2$ where $\sum y_1$ and $\sum y_2$ represent the total numbers of observations in the first and second groups respectively. It may be shown by substitution in this equation that, if $|\bar{X}|=|\sigma_1|$, $\beta_2=2.5$, whereas, if $|\bar{X}|=|10\sigma_1|$, $\beta_2=1$, approximately. Thus, if the numbers of observations in each of the two sub-groups are the same and the component curves are normal, the value of β_2 for the entire distribution about the mean of the two will, in general, decrease as $|\bar{X}|$ becomes large in comparison with $|\sigma_1|$. Differences in β_2 of this magnitude are difficult to establish. Furthermore the skewness is zero, and therefore does not indicate the bi-modal character of the distribution.

Let us consider the case where $|a\,\bar{X}_1|=|\bar{X}_2|$; $k_1=k_2=0$; ${}_1\beta_2={}_2\beta_2=3$; $\sum y_1=a\sum y_2$; ${}_1\mu_i={}_2\mu_i$. If, $a=10$ and $|\bar{X}_1|=|\sigma_1|$ then $\beta_2=8+$ whereas if $|\bar{X}_1|=|10\sigma_1|$, then $\beta_2=100$, approximately.[39] Thus, for comparatively wide differences in the averages, it requires a large number of observations in order to increase the precision of β_2 to such an extent as to prove the significance of deviations in this factor of the magnitudes noted above.

The skewness in this case is not zero and its significance could be established with a comparatively small number of measurements. In any of the above cases a carefully constructed plot would serve to indicate the bimodal characteristic of the curve better than the study of the factor β_2.

Pearson's Criterion of Goodness of Fit. A much more powerful

[39] Here again it should be noted that the values of β_2 are independent of the actual frequencies of each of the two groups and depend only upon the ratio of these frequencies and upon the ratio of $|\bar{X}_1|$ to σ_1.

criterion has been developed by Prof. Pearson [40] in a series of articles in the *Philosophical Magazine.* It is true that this test for goodness of fit cannot be used indiscriminately. In fact the application of this criterion is subject to numerous limitations clearly set forth in the original papers by Pearson and in more recent articles on the mathematics of statistics. In the use of the method it is necessary that these be kept in mind by the individual making the original analysis of the data. Irrespective of these facts, however, the method itself is one of the most useful tools available for measuring in a quantitative way the "goodness of fit" between two distributions. The significance of the values of P given in Figs. 5, 6, and 8 now become evident.

Engineering Judgment. The fourth very practical and one of the most useful methods of comparing the theoretical with the observed distribution is that of applying common sense or engineering judgment. To quote from a recent article of Prof. Wilson [41] we have: "And as the use of the statistical method spreads we must and shall appreciate the fact that it, like other methods, is not a substitute for, but a humble aid to the formation of a scientific judgment." Even with the use of all the statistical methods known to the art, it remains impossible to determine the true nature of the complex of causes which control a set of observations. We can present plausible explanations, but we can never be sure that they are right. Sometimes we can present two plausible explanations and then we must fall back on engineering judgment or common sense to decide between them. A striking illustration of this fact is presented in the following paragraph.

Prof. Pearson [42] has recently presented measurements of the cephalic index of a certain group of skulls. The object of the investigation was to determine if variation had gone on to such an extent as to indicate the survival of the fitter inside a homogeneous population, or the survival of two races both of which were in existence many ages in the past. Pearson shows that, by a solution of a nonic equation,

[40] If we divide the entire range of variation into s equal intervals for which the observed frequencies are $f_1, f_2, \ldots . f_s$ and the corresponding theoretical frequencies are $f_1', f_2', \ldots \ldots . f_s'$, Pearson calculates the function

$$\chi^2 = \sum \frac{(f'-f)^2}{f'}$$

from which he is able to determine the probability that a series of deviations as large as, or larger than, that found to exist could have arisen as a result of random sampling. Tables have been prepared which give the probability of fit in terms of the number of intervals into which the entire range has been divided and of the value of χ.

[41] Wilson, E. B.—The Statistical Significance of Experimental Data—science—New Series, Vol. 58, 1493, October 10, 1923, pp. 93-100.

[42] *Philosophical Magazine,* Vol. 1, 1901—pp. 110-124.

he is able to find two component distributions which when added together approximate very closely to the observed frequencies. The observed data are given in the second column of Table VI and the frequencies of Prof Pearson's compound curve are given in the third column of the table. The probability of fit between these two distributions is seen to be approximately .96, which is indeed very

TABLE VI

ROWGRAVE SKULLS *

Cephalic Index	Observed Distribution f	Compound Distribution f_1	2nd Approximation f_2	$\frac{(f_1-f)^2}{f_1}$	$\frac{(f_2-f)^2}{f_2}$
67	1	1	1	0	0
68	1	2	2	.50	.50
69	3	4	4	.25	.25
70	8	7	8	.14	0
71	13	11	14	.36	.07
72	13	18	22	1.39	3.68
73	33	28	30	.89	.30
74	36	39	39	.23	.23
75	49	50	48	.02	.02
76	59	59	55	0	.29
77	69	65	59	.25	1.69
78	70	66	60	.24	1.67
79	54	60	58	.60	.28
80	58	52	53	.69	.47
81	40	43	46	.21	.78
82	31	35	39	.46	1.64
83	25	28	32	.32	1.53
84	28	23	26	1.09	.15
85	21	20	21	.05	0
86	20	17	16	.53	1.00
87	9	14	13	1.79	1.23
88	10	11	10	.09	0
89	6	8	7	.50	.14
90	10	6	5	2.67	5.00
91	2	4	3	1.00	.33
92	3	2	2	.50	.50
93	2	1	1	1.00	1.00
94	1	1	1	0	0
95	0	0	1		1.00
Σ	675	675	676	15.77	23.75
Probability of fit P				.957	.694

Ave. $=\bar{x}=78.846$ Ave. $=\beta_2=3.181$ Ave. $=\sigma_k=$ .0943

$\sigma=$ 4.612 $\sigma_{\bar{x}}=$ .178 $\sigma_{\beta_2}=$ .189

$k=$ .521 $\sigma_{\sigma}=$ .126

* Phil. Mag., Vol. I, 1901, pp. 115–119.

high, meaning, of course, that 96 times out of 100 we may expect to find a system of deviations as large or larger than that actually found. The author finds, however, that the theoretical distribution (column 4) based upon the assumption of the second approximation is also a very close fit to the observed frequencies, the probability of fit being in this case .69. As a result of these calculations shall we conclude that the distribution is composed of two normal components as indicated in Fig. 10, or shall we conclude that the distribution is homogeneous? In other words, do the skulls belong to two or to only

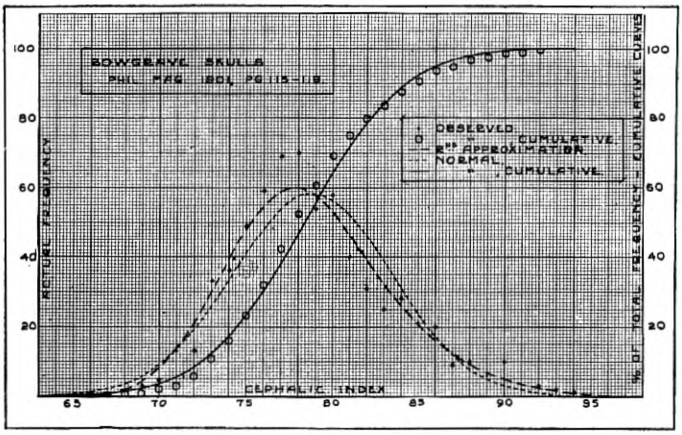


Fig. 10

one race? The measure given by the probability of fit is, of course, in favor of the first alternative. It is highly probable, however, that if we had been given the observed distribution without any discussion of what it meant we would have decided that it probably was consistent with the assumption of the random system of causes such as might underlie the second approximation.

In other words, if we had been given merely the above set of skull measurements, it is reasonable to suppose we might have concluded that the distribution was homogeneous. However, when our judgment is colored by the facts which cannot be presented in the array of observed frequencies we must conclude that it is highly probable that the observed data have arisen from a non-homogeneous population.

Statistical methods alone do not answer all of the questions that are raised in this problem nor do they answer them in many others. There is almost always room for judgment to enter.

Thus, analyzing a group of measurements of some characteristic of a large number of transmitters, it often becomes necessary to determine whether or not they can be subdivided into normal components as in the above problem. In our case the subgroups correspond to different kinds of carbon. Here, as in the data given by Pearson, it often has been found necessary to base our final conclusion partly upon facts not revealed by the data themselves.

The integral curves corresponding to the normal and observed distributions are given in Fig. 10 in order to show that they do not

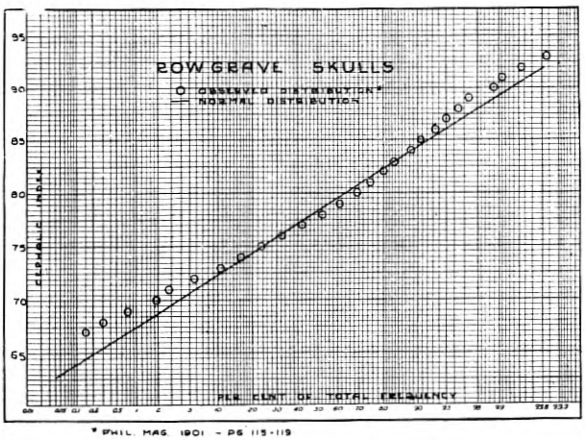


Fig. 11

serve to indicate the difference between the observed and theoretical distributions nearly as well as the actual frequency curves also given in this figure. Fig. 11 presents the result on probability paper. In this case the probability curves are as good as the frequency curves for showing the divergence between theory and observation. It will be recalled that this is not true for the similar curves given in Fig. 9.

Summary Statement of Suggested Method to be Followed in the Analysis of Engineering and Physical Data

We have briefly reviewed the different methods for determining the best theoretical distribution to represent observed data. The following four steps indicate the ordinary procedure:

1. Obtain the first four corrected moments.
2. Calculate the average, standard deviation, k and β_2, and their standard deviations.
3. Calculate the theoretical distribution of distributions warranted by the circumstances.

4. Apply one or more of the four methods of comparing the theoretical and observed distributions to determine which one is theoretically the best.[43]

An illustration of the method of applying this form of analysis to inspection data on transmitters is indicated in the schematic chart Fig. 12. The object of the inspection of apparatus in the process

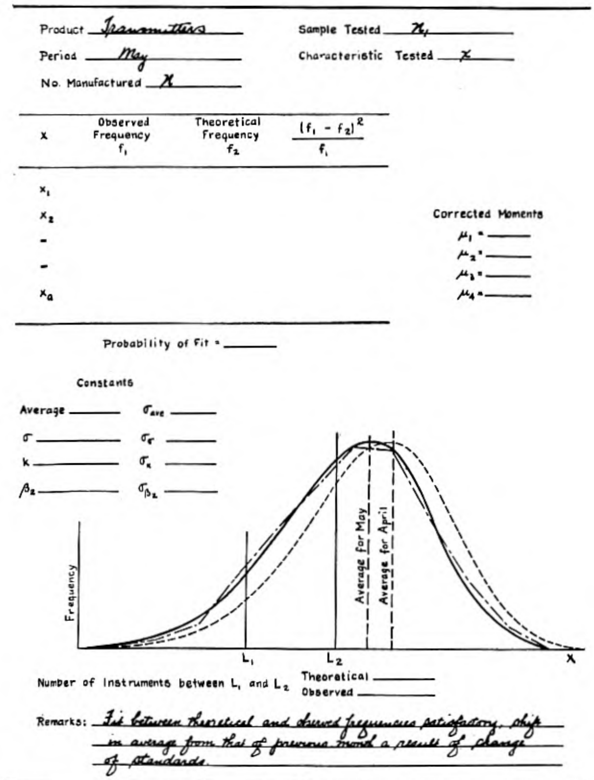


Fig. 12

of manufacture is obviously to determine the most probable law of distribution, and from this to determine whether or not there is any indication of a trend in the quality of the product. In the light of what has been said, it is obvious that a complete report of this character should contain the items called for in Fig. 12. The corrected

[43] If the observed distribution could not have arisen from a random system of causes, it may be advisable to attempt to transform it into an approximately random one, such as was done in connection with the data in Fig. 6.

moments and the factors, such as the average, standard deviation, k and β_2 should be given. These factors provide us with measures of the lack of symmetry, and can be used as pointed out in the previous sections of this paper. Recording this amount of data makes it possible for anyone interested, either to check the calculations of the theoretical frequencies and the conclusions derived therefrom, or to calculate a different theoretical distribution based upon fundamentally different hypotheses in a way such as has been illustrated already in the discussion of the distribution of measurements of the cephalic index, as given in Fig. 11.

In most instances, however, it is highly probable that the man who originally prepares the chart is charged with the responsibility of choosing the best distribution, and, therefore, the chief interest of those reading the report is centered upon the conclusions indicated therein. The graphical representation of the observed distribution by means of the histogram is hopeful. The comparison of this with the theoretical curve represented by a solid line shows qualitatively whether or not the product is changing. The probability of fit gives a quantitative measure of the degree of fit. The set of curves given in Fig. 12 is drawn to illustrate a condition which may sometimes happen when, for example, the standards used in the machines have been changed. This is only typical of the results which may be expected. Obviously, the form of such reports designed to meet specific conditions will vary. That presented above is only typical of one which has been found to be of value in presenting the analysis of the results of inspection of certain types of apparatus.

Some Advantages Derived from a Comparatively Complete Statistical Analysis

It has been pointed out that the value of either a physical or an engineering interpretation of data depends upon the success attained in deriving the best theoretical distribution. This is the equation which fits the observed points best, and which, if possible, can be interpreted physically. The previous discussion indicates the way in which different causal relationships tend to produce typical frequency distributions, and also the way in which statistical methods may be used in finding a theoretical distribution which yields a physical interpretation.

This point has been illustrated by several examples. It has been shown that by a proper choice of theoretical curve a very close approximation to an observed distribution can be obtained. This

TABLE VII—FREQUENCY DISTRIBUTION OF σ PARTICLES *

Number of Particles	Observed Frequency f	†Normal Law f_1	2nd Approximation f_2	‡Law of Small Numbers f_3	§Poisson Charlier f_4	$(p+q)^n$ f_5	$\frac{(f-f_1)^2}{f_1}$	$\frac{(f-f_2)^2}{f_2}$	$\frac{(f-f_3)^2}{f_3}$	$\frac{(f-f_4)^2}{f_4}$	$\frac{(f-f_5)^2}{f_5}$
−3	0	1	0	0	0		1.000				
−2	0	5	0	0	0		5.000				
−1	0	23	8	0	0		23.000	8.000			
0	57	74	63	54	47	50	3.905	.571	.167	2.128	.180
1	203	180	197	210	196	202	2.939	.183	.233	.250	.005
2	383	337	389	407	400	405	6.279	.093	1.415	.123	1.195
3	525	485	532	525	533	533	3.299	.092	0	.120	.120
4	532	535	526	508	525	521	.017	.068	1.134	.093	.232
5	408	452	398	394	407	402	4.283	.251	.497	.002	.090
6	273	294	249	254	258	255	1.500	2.313	1.421	.872	1.271
7	139	146	139	140	138	137	.336	0	.007	.007	.027
8	45	57	72	68	64	64	2.526	10.125	7.779	5.641	5.641
9	27	15	28	29	25	26	9.600	.036	.140	.160	.038
10	10	4	10	11	9	9	9.000	0	.091	.111	.111
11	4	1	3	4	3	3	9.000	.333	0	.333	.333
12	**0	0	1	1	1	1	∞	1.000	1.000	1.000	1.000
13	**1	0	0	0	0	0					
14	**1	0	0	0	0	0					
Σ	2608	2609	2615	2605	2606	2608	81.684	23.065	13.884	10.615	10.245
Probability of Fit							.0000	.0410	.3086	.5624	.5946

$p = .046$ $\qquad \beta_2 = 3 + \frac{1-6\,pq}{pqn} = 3.51 \qquad \sigma_{\beta_2} = \sqrt{\frac{24}{2608}} = .0959$

$q = .954$

$\bar{x} = pn = 3.87$

$\sigma = \sqrt{pqn} = 1.92 \qquad \sigma_k = \sqrt{\frac{6}{2608}} = .0480$

$K = \frac{q-p}{\sqrt{pqn}} = .48$

$n = 84.174 \qquad$ ($N=64$ used in computing $(p+q)^n$)

* Prof. Rutherford and H. Geiger—Phil. Mag., Oct., 1910.

† Normal calculated on basis of probability integral.

‡ Bateman, H., in appendix to original article.

§ Fisher, A.—loc. cit., p. 273,

** These observed frequencies are grouped together in computing P.

has already been indicated in Table III. To emphasize this point, however, let us consider once more the distribution of alpha particles given in Table I. These data together with various theoretical [44] distributions are given in Table VII.

Let us consider the data given in Table I by following the procedure of analysis outlined in the previous section. The factors k and β_2 when compared with their errors should indicate whether or not the distribution is normal. As shown in Table VII, k and β_2 differ from 0 and 3 respectively, by more than 3 times their respective standard deviations. As has already been pointed out, this is sufficient evidence to indicate that the distribution is not normal. In order to show, however, that if we follow the next step and calculate theoretical distributions based upon the assumption of the different laws; that is, in this case, normal, second approximation, and the law of small numbers, we are naturally led to the choice of the best distribution. This choice is materially influenced by the measure of the probability of fit as recorded in the table. The law of small numbers is obviously a very close approximation to the observed frequencies.

One of the obvious things to do in this problem, but one that has not been done previously, is to calculate the values of p, q and n, and from them the terms of the binomial expansion $2608(p+q)^n$. The probability of fit between the terms of this expansion and the observed frequencies is the highest given in the table. This increases the evidence that the distribution is random. It also does nore. It serves to establish the facts that the probability p that an alpha particle will strike the screen is .046, and that the maximum number of alpha particles which may ever be expected to strike the screen is of the order of magnitude of 84. Granted then that we can always find the most probable theoretical frequency distribution, let us consider next the influence that the result may have in our determination of the most probable value, the number of observations between any two limits and the casual relationships governing the distribution.

Let us consider first the dependence of the most probable value upon the type of distribution. In our present work in the study of carbon the resultant distributions have been in most instances either random or such that through a proper transformation they could be reduced to such. For any distribution consistent with the second approxima-

[44] The source of all distributions previously calculated are indicated. The Poisson-Charlier series is similar to the Gram-Charlier series, except that the law of small numbers is the generating function. It serves as an admirable method of graduating certain classes of skew distribution as illustrated by this example and by that given in Table III.

tion the most probable value is at a distance $-\frac{k\sigma}{2}$ from the arithmetic mean. Many distributions have been found for which k lies between .5 and unity, and, therefore, this difference is from ¼ to ½ of the standard deviation. Thus, the efficiencies of certain standard types of transmitters are found to conform to such a law, and the difference between the modal and average values is of the order of magnitude of 0.4 mile.

Obviously the geometric mean of the sound intensities (Fig. 6) and not their arithmetic mean is the most probable. The difference between the two is quite large. The difference between the arithmetic mean and the modal value for groups of data such as given in Fig. 1, Tables II and VI are quite large. To use again the illustration ot the alpha particles the observed most probable number is 4; whereas, the observed average [45] is 3.87. Judging from the best theoretical distribution the most probable number of alpha particles is 3. Choosing the number 3 it is seen that either of the other two numbers differ from this by approximately ½ the standard deviation. Such results are, however, not confined to the work of the present investigation nor to the examples previously cited as is evidenced by the data given in the last column of Table VIII.

TABLE VIII

N = Number of Observations	Source of Data	Percentage Within $\overline{X} \pm \sigma$	Percentage Within $\overline{X} \pm 2\sigma$	Percentage Within $\overline{X} \pm 3\sigma$	$\frac{\text{Average} - \text{Modal}}{\sigma}$
1000	E *54	66.6	97.2	99.6	.803
251	E 66	78.1	94.8	97.6	1.042
9154	E 10	67.7	95.5	99.6	.031
2162	E 79	70.1	95.1	99.3	−.311
368	E 84	73.4	94.6	97.0	.422
675 Table VI.........		68.7	94.1	99.6	.247
Normal Law......		64.26	95.44	99.73	0

* Elderton "Frequency Curves and Correlation," published by C. & E. Layton, London, 1906.

We should not leave this phase of the discussion, however, without pointing out that in a large number of purely physical experiments a sufficient number of observations has not been taken to make it possible to choose the best theoretical distribution. In general more than

[45] Of course, such an average has no significance, except for a continuous distribution.

100 observations are required. Thus, in Prof. Millikan's [46] determination of the electron charge e only 58 observations were made. The values of σ, k, and β_2 for this distribution are .128 units, $-.196$ and 2.358. Even though the observed distribution is consistent with a normal system of causes, values of k and β_2 may be expected to occur which differ from 0 and 3 respectively, as much as these observed values do. In this case even if k is real and not a result of random sampling, the correction to be added to the average in order to obtain the most probable value is insignificantly small.

Next let us consider the problem of determining the number of observations between any two limits. The physicist is ordinarily concerned with the probable error: that is, the error such that ½ of the observations lie within the range $\overline{X}\pm$ probable error. Its magnitude for the normal distribution is $.6745\sigma$, and the errors are distributed symmetrically on either side of the average. It is interesting to note that the magnitude of the probable error is also $.6745\sigma$ for the second approximation, but that the errors are not distributed symmetrically on either side of the average.

Another important pair of limits is that including the majority of the observations. For the normal law 99.73% of the observations are included within the range $\overline{X}\pm 3\sigma$ which, therefore, is often called the range. Not a single example has been found, however, of a distribution for which the observed number of observations within this range is less than 95% even though the distribution is decidedly skew. In fact it is seldom less than 98%. If, however, we have a case such as that represented in Table II where groups of observations have been taken in what is technically known as different universes, and then averaged together, the average result is not the most probable, and the standard deviation of the average is not inversely proportional to the square root of the number of observations. Since this point is of considerable importance, it is perhaps well to state it in a slightly different way. Thus, let us assume that we have a thousand samples of granular carbon which possess inherent microphonic efficiencies differing by comparatively large magnitude. Transmitters assembled from any one of the groups of carbon cover a range of efficiencies. If we choose a sample of 10,000 instruments, 5,000 from each of two lots of carbon which do not possess the same inherent efficiency, we cannot expect, for reasons already pointed out, that the observed distribution will be normal. The average of these observations will not in general be the most probable value, and the standard deviation of the average will not be equal to the

[46] Millikan, R. A.—The Electron—*University of Chicago Press.*

observed standard deviation divided by the square root of the number of observations, in this case 10,000.

We have already seen, however, that it is possible to detect such errors of sampling, since in general the distribution cannot be fitted by the second approximation or Gram-Charlier series. If the theoretical distribution is either normal, second approximation, or the law of small numbers, the number of observations to be expected between any two limits can be readily determined from the tables. Experience has shown that in every instance where it has been possible to represent the observed distribution in any of these three ways, the data obtained in future samplings have always been consistent with the results to be expected from the theory underlying these three laws. It will be of interest to note the data given in columns 3, 4, and 5 of Table VIII and to compare the theoretical percentages (last row) for the different limits with those observed.

In closing it is of interest to point out further the significance of some of the results discussed in this paper in connection with the inspection of equipment. Here we must decide upon a magnitude of the sample to be measured in order to determine the true percentage of defective instruments in the product. If p is the percentage defective, and q that not defective, then the standard deviation about the average number found in a sample of n chosen from N instruments is

$$\sigma^2 = pqn\left(1-\frac{n}{N}\right).$$

In practice, however, we never know the true value of p unless we measure all of the apparatus, and this is impractical. In our calculations we must therefore use some corrected value. We find, though, that the average value of p is in most instances the one that must be used. Assuming that we choose a value of p, the distribution of defectives in N' samples of n in number will be represented by the distribution of $N'(p+q)^n$. If one of the samples is found to contain a percentage of defectives, which is inconsistent, that is, which is highly improbable as determined from the distribution of $N'(p+q)^n$, it indicates that the product is changing.

If, however, we take into account the effect of the size of the first sample in respect to the second as indicated by Pearson,[47] we see that the distribution of N' samples may be different from that given by the binomial expansion. In accordance with this theory, if in a first sample of 100, 10% of the sample is found to possess a given attribute,

[47] Pearson, K. Loc. cit. Foot note 30.

the distribution of the percentages to be expected in 1,000 such samples is indicated by the last column of frequencies in Table III. In order to show graphically how this distribution differs from that

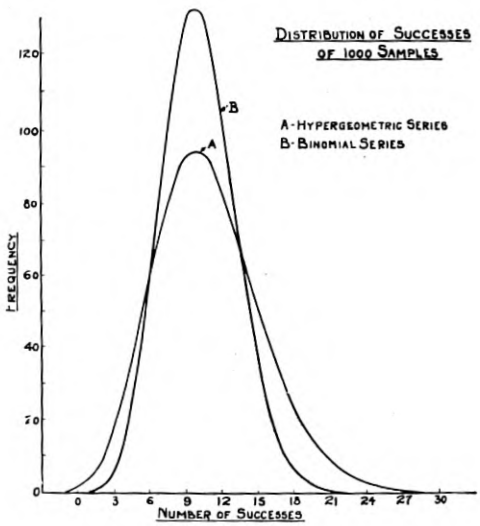


Fig. 13

corresponding to the binomial expansion these two sets of frequencies are reproduced in Fig. 13. The difference between them is a striking illustration of the significance of the size of the samples used in connection with the inspection of equipment, providing we accept Pearson's results.

Deviation of Random Samples from Average Conditions and Significance to Traffic Men

By E. C. MOLINA and R. P. CROWELL

THE traffic executive deals with questions which lead him into the consideration of problems of widely differing natures. At almost every turn he is confronted by the fact that his decisions and programs in relation to these different phases of the work must be based on records which are seldom continuous and in most cases are merely "samples." These sample records are assumed to measure the characteristics of the entire volume of facts or data of which they are taken to be representative. In the use and analysis of these records there are a number of perplexing questions which come to his mind if he allows himself the luxury of a little theoretical speculation.

Practically all of his information regarding the efficiency with which his office is run and on which he must base his plans for continued efficiency is obtained from the peg counts. These peg counts are records of the number of calls handled and are taken on two or three days out of each month. At the same time that the calls are counted, the number of employee hours used in the handling of the traffic is counted. The results of these peg counts are used to represent the performance of that office for the month. When the inquiring traffic man meditates a little on the subject of these peg counts he soon begins to wonder how nearly representative they are of his every day performance. He can—and sometimes does—think up a number of things which will explain any poor results which show up.

One of the means taken to insure the accuracy of the peg count is to observe the counting of 25 to 50 calls each by as many of the operators as possible, with the idea of determining how accurately the operators count. In this way from 1,500 to 3,000 observations are made on the accuracy of the operators' counting, in a period of two or three days. The traffic man occasionally questions whether he can rely on the results of this comparatively small number of checking observations to give him an indication of the accuracy of the count as a whole.

In order that comparisons may be made of the performance of different offices and the cost of handling different kinds of calls, it is the practice to translate all the work done into terms of traffic units (representing the relation of the labor value of the different operations to a fixed value arbitrarily selected). In order to do this, at longer intervals than the regular peg counts, the traffic is counted in more detail. From certain classifications and subdivisions of these supplementary counts, coefficients or equating factors are developed which are applied to the regular counts to develop units. The speculative traffic man ponders over these and wonders how representative the supplementary counts are of the every day distribution of traffic.

This speculation leads him also to question the labor values which have been assigned to the different operations and which have been furnished him for the purpose of equating his traffic. He knows that because of the impossibility of making continuous stop watch observations on his operators, he has to accept the results of such observations made on a considerable number of calls handled in a similar manner at some time in the past and probably in some other place, as being representative of the work involved in handling those types of calls at the present time in his office.

After thus puzzling himself over peg counts and similar records, the traffic man may turn his attention to some of the service problems and begins to scrutinize with considerable skepticism the records which are maintained of this feature of his work. Among the most valuable records of the way in which the service at his office is being handled, are the records developed as a part of the central office instruction routines. These are observations taken on ten calls handled by each of the operators on the force, periodically. He looks over the latest detail sheets and observes that the results of these tests on two particular operators show that the one he considered a very careful and methodical girl has made a high proportion of mistakes while the operator whom he thinks is the more careless shows an absolutely perfect test. Because of his other knowledge he suspects these records and decides to check them up by examining the summaries of similar tests taken for some months past. These summaries show figures which bear out his original estimate of the ability of the two operators, which relieves his mind but leaves him still puzzled as to why the averaging of a series of figures which are not representative, makes the summary more nearly representative.

There is another set of figures which the traffic man consults in connection with the quality of the service and which causes him a good deal of worry. These are the figures obtained from central office speed of answer tests, tests of the speed of answer to recall signals, etc. The speed of answer tests, for example, are made by an employee in the central office who causes signals to appear and with a stop watch determines how long it takes the operators to answer each signal. The signals used in making these tests are distributed in all parts of the switchboard and the number of tests made in each hour is roughly proportional to the amount of traffic handled. The results of these tests are summarized in such a manner as to show the percentage of tests which are not answered within 5, 10 and 20 seconds. The traffic man who gives this matter thought, is concerned to know how much reliance he can place on the results of these tests as being representative of the percentage of slow answers applying to all the calls handled in the office.

The speculative traffic man by this time is in a frame of mind which either leads him to doubt all figures or to feel that there must be something in the figures which he cannot explain but which makes certain of them quite representative, although there are certain others about which he does not feel the same way. He is sure that some of them are representative because decisions and programs based on them produce the results desired. He is also sure that some of them are not representative because they imply things which he knows are not so, as a result of observation. Just how far he can rely upon the figures which he is using, and where to draw the line is a question which only long experience or an understanding of the reasons which lie behind the taking of these records can solve. It will probably be of interest to discuss, from the purely theoretical angle, certain simple traffic data with the idea of noticing how the application of a certain mathematical procedure can aid in drawing accurate conclusions from them.

The type of traffic problem which will be considered may be stated as follows:

A group of 50,000 calls originated in an exchange area. An unknown number of them were delayed more than 10 seconds. Observations were made on 300 of the calls and of these 9, or 3 per cent., were delayed more than 10 seconds. With this information is it a safe bet that the unknown percentage for the entire 50,000 calls is below 5? Or better yet, are we justified in betting 99 in 100 that the unknown percentage for the 50,000 calls is below 5? Or again, may we bet 8 in 10 that the unknown percentage is between 0.5 and 5? It is taken for granted that the observer is justified in believing that the calls under consideration fulfill the conditions of random sampling such as that each call is independent of every other call, or that an appreciable number of the calls is not due to the occurrence of some unusual event,—the opening of the first game of the world series, for example.

Assuming that the reader is unfamiliar with the theory of probability, a digression becomes necessary and in order that he may enter into the spirit of the theory the reader is requested to forget for the present the telephone problem. Of course, only a bird's-eye view of the theory will be given here. Several lacunæ will be encountered; the filling in of any one of them would call for a volume of not very small dimensions.

INTRODUCTION TO THE THEORY OF "A POSTERIORI" PROBABILITY

The problem to be dealt with belongs to the class of problems which gave rise to that branch of the Theory of Probability which is known as "A Posteriori Probability" or "Probability of Causes." It is frequently referred to as the Theory of Sampling.

To bring out certain of the ideas involved it will be helpful to consider what may appear as a very extreme example from the traffic man's point of view, but which is nevertheless typical of the type of problem in which a consideration of a posteriori probability enters. We are told that at a student gathering a particular young man won 7 out of 15 times. Our informant refuses to divulge what is going on at the gathering. What probabilities should we assign to the following hypotheses?

1. He threw heads 7 times out of 15 throws with a coin.
2. He threw 7 aces out of 15 throws with a 6 face die.
3. He won on points 7 rounds in a fifteen round bout.
4. The aggregate of all other hypotheses.

A little careful consideration will make it clear that with reference to each hypothesis (or aggregate of hypotheses) two essential questions must be answered before we can determine the a posteriori probability. Consider the six face die hypothesis; we must know:

1st —What is the relative frequency or probability with which gambling with a 6 face die is indulged in at student gatherings?

2nd—Given a six face die, what is the probability of throwing an ace 7 times in 15 throws?

Quoting Mr. Arne Fisher[1] we may restate these two questions as follows:

1st —What is the a priori *existence* probability in favor of the 6 face die hypothesis?

2nd—What is the *productive* probability for the observed event given by the hypothesis of a 6 face die?

[1] Arne Fisher—The Mathematical Theory of Probabilities—2nd Edition—Art. 41.

In most problems of this type the determination of the *productive* probability for each hypothesis is a question of pure mathematics. But when we proceed to evaluate the a priori *existence* probability for each hypothesis or cause, common sense and guessing must frequently be resorted to. The history of the applications of a posteriori probability is so full of paradoxes resulting from appeals to common sense that to some high authorities the whole theory is a fallacy. Prof. George Chrystal [2] closes a severe attack on Laplace's Theorie Analytique with the statement—"The indiscretions of great men should be quietly allowed to be forgotten." Nevertheless, the writers will assume the Laplacian view of the subject, especially as it has been defended by such authorities as Karl Pearson and E. T. Whittaker.

The above typical problem has been introduced because its mere statement leads us immediately to the conceptions of existence and productive probabilities with reference to different possible hypotheses. But, it is not our intention to bring any notoriety on the young man by answering the questions raised. Moreover, the hypotheses made, differ qualitatively, whereas, our telephone problem involves various hypotheses which differ only quantitatively. We, therefore, proceed to another typical problem, a solution of which will give us at once the solution of the telephone problem.

A bag contains 1,000 balls; an unknown number of these are white and the rest not white. Of 100 balls drawn 7 are found to be white. What light does this information throw on the value of the unknown number of white balls? What is the probability that there are 70 white? Is it a safe bet that the number of white balls lies between 60 and 80?

Two cases of this problem may be considered:

Case 1. After a ball is drawn it is replaced and the bag is shaken thoroughly before the next drawing is made.

Case 2. A drawn ball is not replaced before another ball is drawn.

These two cases become essentially identical if the total number of balls in the bag is very large compared with the number drawn.[3] In the following discussion Case 1 is assumed.

The information at hand is that 100 drawings resulted in 7 whites. Obviously the bag contains at least one white, but we are free to choose between 999 possible hypotheses.

[2] Transactions of the Acturial Society of Edinburgh—Vol. II, No. 13—On Some Fundamental Principles in the Theory of Probabilities.

[3] For the application to practice herein contemplated it is thought that the number of balls in the bag should be at least ten times the number drawn.

1—The bag contains 1 white and 999 not white.
2—The bag contains 2 white and 998 not white.
3—The bag contains 3 white and 997 not white.
....
K—The bag contains K white and (1,000-K) not white.
....
997—The bag contains 997 white and 3 not white.
998—The bag contains 998 white and 2 not white.
999—The bag contains 999 white and 1 not white.

Let $W(K)$ be the existence probability for the K'th hypothesis. By "existence probability" is meant the likelihood that the bag contains exactly K white balls when the circumstances of the drawing, but not the actual results of the drawing, are fully taken into account. Its exact value may often be in doubt either because we do not have complete knowledge of the circumstances preceding the drawing or because we are not able to deduce its exact value from this knowledge. It is obvious, however, that there must be some such value and we must, therefore, introduce a symbol to represent it.

Let $B(7,100,K)$ = productive probability for the K'th hypothesis; by this is meant the probability of obtaining the observed event (7 white in 100 drawings) if the bag contains K white balls and 1,000-K that are not white.

Then the a posteriori probability in favor of the K'th hypothesis (meaning thereby the probability in favor of the K'th hypothesis *after* the 7 white balls were drawn) is [4]

$$P_k = \frac{W(K)B(7,100,\, K)}{\sum_{s=1}^{s=999} W(S)B(7,100,\, S)}. \tag{1}$$

Now to say that the bag with a total of 1,000 balls contains K white balls is equivalent to saying that the *ratio* of white to total balls is

$$p_k = K/1000$$

and that the *ratio* of not white to total balls is

$$q_k = 1 - p_k = (1000 - K)/1000.$$

[4] This is the celebrated Laplacian generalization of Bayes' formula. No attempt to demonstrate it will be made here. The subject is dealt with at length by Laplace in the Théorie Analytique des Probabilités and by Poisson in the Recherches Sur La Probabilité des Jugements. A beautiful and relatively short demonstration is given by Poincaré in his Calcul des Probabilités.

We may, therefore, rewrite (1) as follows:

$$P_k = \frac{W'(p_k)B'(7{,}100,\ p_k)}{\sum\limits_{s=1}^{s=999} W'(p_s)B'(7{,}100,\ p_s)}, \tag{2}$$

where W', B' are the forms assumed by the functions W, B, respectively, when the ratio p_k is used instead of the number K.

The interpretation of the terms of the expansion of the binomial $(p+q)^{100}$ tells us that

$$B'(7{,}100,\ p) = \binom{100}{7} p^7 (1-\mathrm{p})^{93} = \binom{100}{7} p^7 q^{93}$$

where $\binom{100}{7}$ is a symbol for the number of combinations of 100 things 7 at a time.

Substituting in (2) and canceling from numerator and denominator the common factor $\binom{100}{7}$ gives

$$P_k = \frac{W'(p_k)p_k{}^7(1-p_k)^{93}}{\sum\limits_{1}^{999} W'(p_s)p_s{}^7(1-p_s)^{93}}. \tag{3}$$

From (3) we obtain for the a posteriori probability that the ratio of white balls does not exceed $K_2/1{,}000$,

$$P(K \ngtr K_2) = \sum_{1}^{K_2} P_k.$$

Likewise, the a posteriori probability that the ratio is not less than $K_1/1{,}000$ is

$$P(K \nless K_1) = \sum_{K_1}^{999} P_k.$$

Finally, the a posteriori probability that the ratio is not less than $K_1/1{,}000$ or greater than $K_2/1{,}000$ is

$$P(K_1 \leq K \leq K_2) = \sum_{K_1}^{K_2} P_k = \frac{\sum\limits_{K_1}^{K_2} W'(p_s)p_s{}^7(1-p_s)^{93}}{\sum\limits_{1}^{999} W'(p_s)p_s{}^7(1-p_s)^{93}}. \tag{4}$$

Solution of the Telephone Problem

Obviously the telephone problem is analogous to the problem of the bag containing an unknown ratio of white balls. The corresponding elements in the two problems may be tabulated as follows:

1st—1,000 balls in bag versus 50,000 calls originated.

2nd—100 balls drawn versus 300 calls observed.

3rd—7 white balls drawn versus 9 calls delayed more than 10 seconds (*i.e.*, defective with reference to a particular characteristic).

4th—To the 999 possible hypotheses with reference to the unknown per cent. of white balls correspond 49,999 possible hypotheses with reference to the unknown per cent. of calls delayed more than 10 seconds.

The problems differ in that a ball drawn from the bag is returned before another drawing is made, whereas an observed call is comparable to a ball being drawn and not returned. With the numbers involved, however, the discrepancy may be ignored.

A formula of the same form as (4) will, therefore, give the answer to our question. We may, however, substitute definite integrals in place of the finite summations since the difference between any two consecutive possible values for the unknown ratio is very small. The integrals together with some desirable transformations of them will be found in the appendix to this article. We will mention here, however, that the transformations made involve an arbitrary assumption as to how the a priori *existence* probability for the different hypotheses varies. As stated above in connection with Prof. Chrystal's views, this is the phase of the subject which lends itself to considerable difference of opinion. The reader who contemplates using the curves embodied in this article should read the appendix with special reference to the assumptions made.

The attached curves Fig. 1 show graphically the conclusions to be drawn from the mathematical analysis. A glance at the right hand end of the curves will show that they are associated in pairs. The upper curve of a pair slopes downward from left to right while its mate slopes upward.

Consider the pair of curves marked .03. For the abscissa 300 they give as ordinates the values .0625 and .014. The interpretation of these figures is as follows: if 300 observations gave 3 per cent. of calls delayed then we may bet

1st—99 in 100 that the unknown percentage of calls delayed is *not greater* than 6.25.

2nd—99 in 100 that it is *not less* than 1.4 per cent.
3rd—98 in 100 that it lies between 1.4 per cent. and 6.25 per cent.

Likewise, considering the curves marked .06 if 1,000 observations gave 6 per cent. of calls delayed, then we may bet

1st—99 in 100 that the unknown percentage of calls delayed is not greater than 8.05.
2nd—99 in 100 that it is not less than 4.4 per cent.
3rd—98 in 100 that it lies between 4.4 per cent. and 8.05 per cent.

It is obvious from the shape of the curves that a few hundred observations do not give more than a vague idea as to the unknown per cent. of calls delayed. On the other hand, the gain in accuracy obtained by making more than 10,000 observations would hardly justify the expense involved. The number of observations which safety requires in any particular problem must be determined by the conditions of the problem itself. If we are willing to take a chance of 9 in 10 or 8 in 10 instead of 99 in 100 or 98 in 100, respectively, the curves of Fig. 2 will give us an idea of the range within which the unknown percentage of defectives lies.

APPENDIX

Case No. 1—Infinite Source of Samples

An inspection of n samples has given c defectives. The observed frequency is then c/n. Let p be the unknown true frequency and p_1 the frequency of delayed calls which has been arbitrarily chosen as being the maximum permissible.

The a posteriori probability that $p \ngtr p_1$ is

$$P = \frac{\int_o^{p_1} W(x) x^c (1-x)^{n-c}\, dx}{\int_o^1 W(x) x^c (1-x)^{n-c}\, dx}, \tag{1}$$

where $W(x)$ is the a priori existence probability that $p = x$. This formula is unmanageable if the form of $W(x)$ is unknown.

Assume first that $W(x)$ is a constant b for $o < x < g$, where $g > p_1$. Then

$$P = \frac{\int_o^{p_1} x^c\ (1-x)^{n-c} dx}{\int_o^g x^c\ (1-x)^{n-c}\, dx + \int_g \frac{W(x)}{b} x^c (1-x)^{n-c} dx}. \tag{2}$$

Now assume that

$$\int_g^1 \frac{W(x)}{b} x^c(1-x)^{n-c}dx,$$

is negligible compared with

$$\int_o^g x^c(1-x)^{n-c}dx,$$

and also assume that g, c and $(n-c)$ are such that approximately

$$\int_o^g x^c(1-x)^{n-c}dx = \int_o^1 x^c(1-x)^{n-c}dx.$$

Then, finally,

$$P = \frac{\int_o^{p_1} x^c(1-x)^{n-c}dx}{\int_o^1 x^c(1-x)^{n-c}dx} = \frac{(n+1)!}{c!(n-c)!}\int_o^{p_1} x^c(1-x)^{n-c}dx, \qquad (3)$$

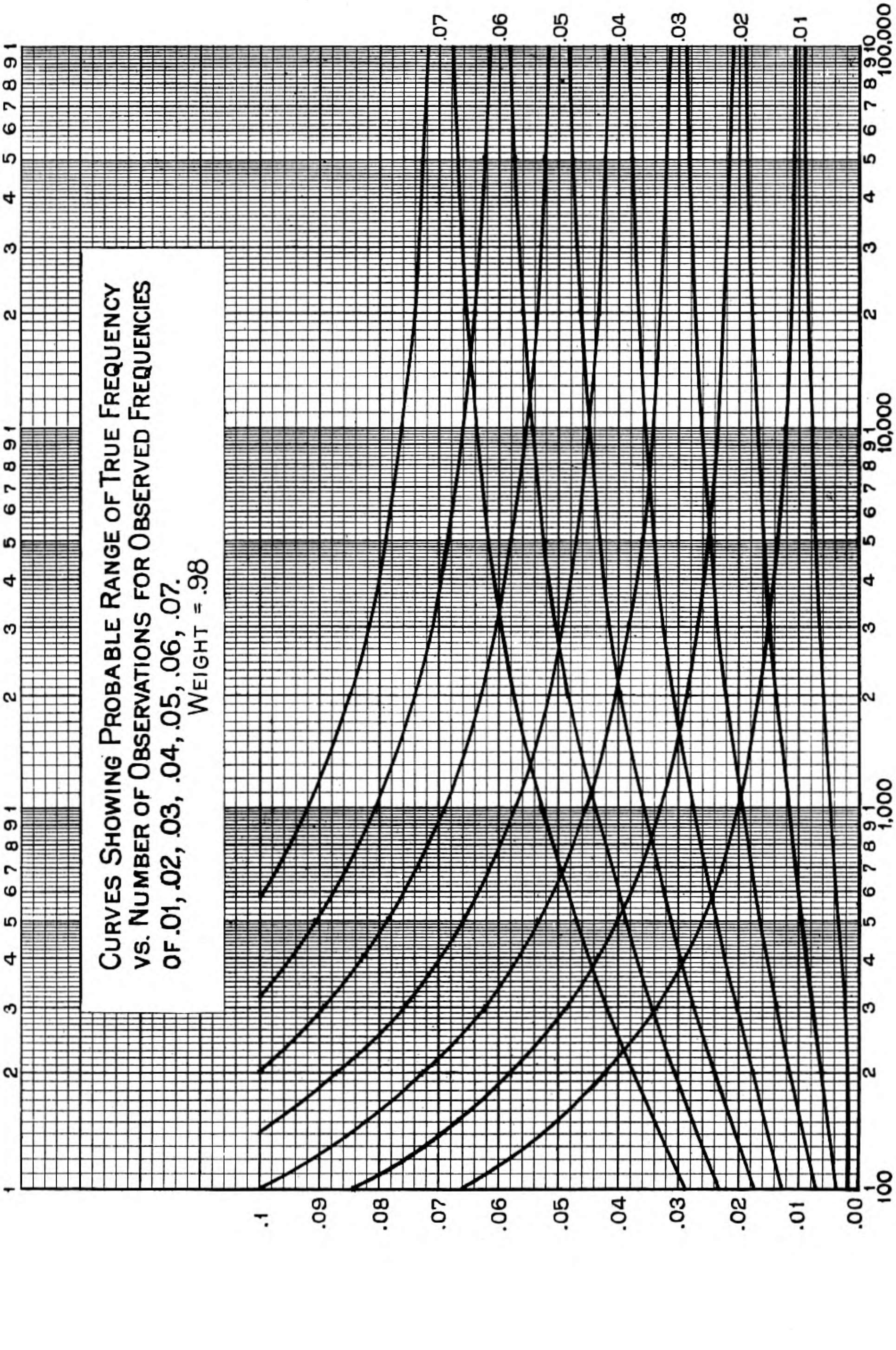
This well known formula might have been obtained by assuming *ab initio* that $W(x)$ is independent of x. It should be particularly noted that this independence is not identical with the assumptions made above. In the applications which are here contemplated the values of p_1, c and n are such that g need be but a small fraction of the range o to 1.

In the "Théorie Analytique" Laplace transforms (3) so that it can be evaluated in terms of the Laplace-Bernoulli integral

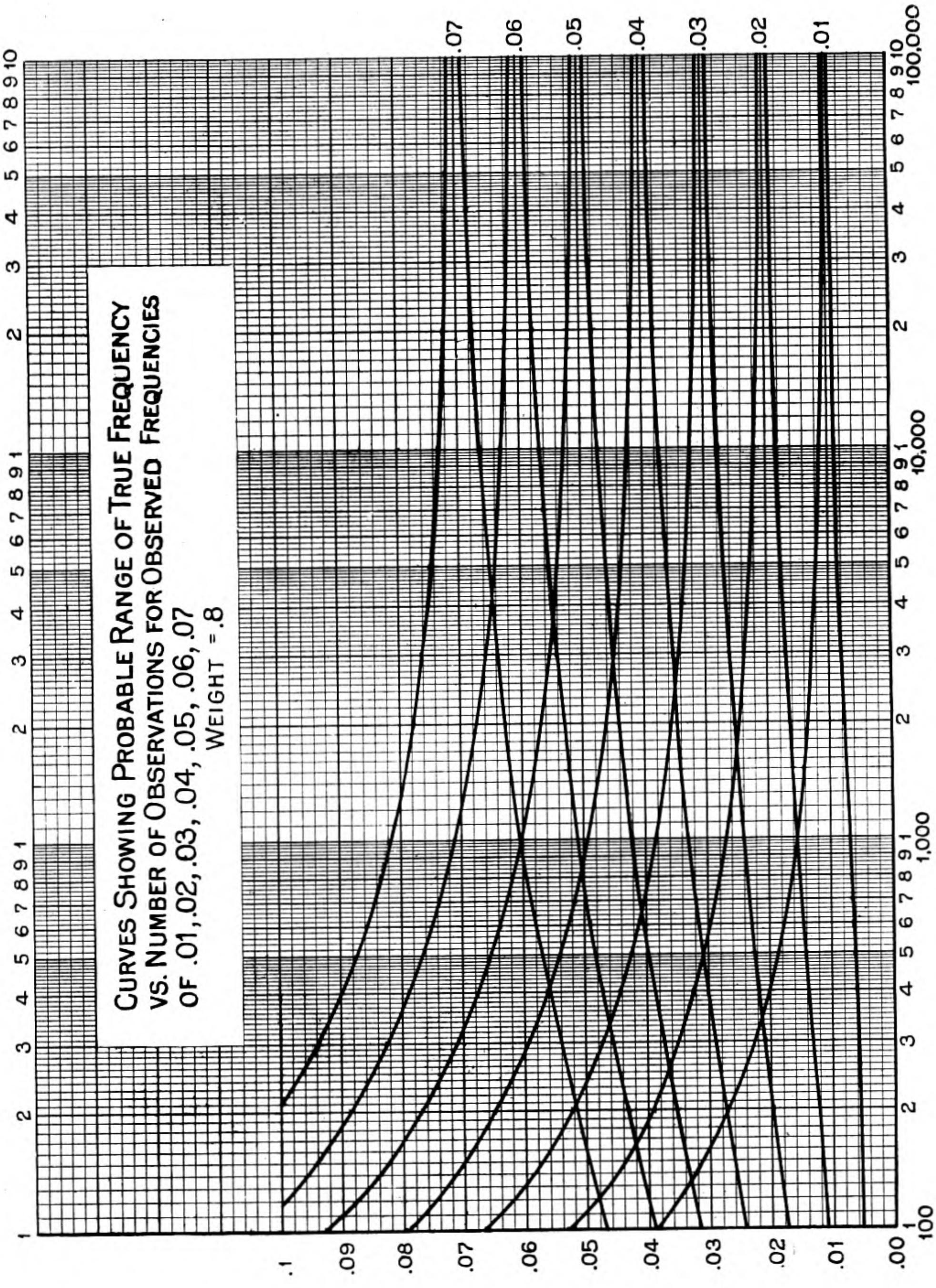
$$\frac{2}{\sqrt{\pi}}\int_o^k e^{-t_2}dt,$$

where k is a function of p_1, c and n. This transformation is most valuable when p_1 is in the neighborhood of 1/2. For small values of p_1 the transformation which converts the binomial expansion to Poisson's exponential binomial limit is more appropriate and gives, writing $(n\, p_1) = a_1$,

$$P = \frac{1}{c!}\int^{a_1} y^c e^{-y} dy = P(c+1, a_1). \qquad (4)$$



CURVES SHOWING PROBABLE RANGE OF TRUE FREQUENCY
VS. NUMBER OF OBSERVATIONS FOR OBSERVED FREQUENCIES
OF .01, .02, .03, .04, .05, .06, .07.
WEIGHT = .98
100
1,000
10,000
100,000
.1
.09
.08
.07
.06
.05
.04
.03
.02
.01
.00


CURVES SHOWING PROBABLE RANGE OF TRUE FREQUENCY
VS. NUMBER OF OBSERVATIONS FOR OBSERVED FREQUENCIES
OF .01, .02, .03, .04, .05, .06, .07
WEIGHT = .8

Photomicrography and Technical Microscopy in Its Application to Telephone Apparatus

By FRANCIS F. LUCAS

NOTE—The following paper may be considered as introductory to the subject of photomicrography. Doubtless everyone is casually familiar with photomicrographs of the crystalline structure of various metals. The application of this branch of the optical art to the study of metals is very important in the design and manufacture of telephone apparatus but its importance in telephony is more far-reaching than in the study of metals alone. Various of these applications are suggested by the illustrations reproduced in the Appendix of this article.—*Editor.*

INTRODUCTION

BY photomicrography is meant the adaptation of photography to microscopy, or the art of photographing a magnified image. The scope of the art embraces the reproduction of images ranging from natural size up to magnifications of several thousand times, the degree of magnification being expressed in terms of diameters. It will be seen that the image is not always magnified but in some instances may be at a 1:1 ratio or when large subjects are being photographed, at an actual reduction in size. Such low-power work is often spoken of as gross photography but so far as the equipment and technique of treatment is concerned it is low-power work.

Low-power work may be considered as treating with magnitudes from 1 to about 30 diameters. Medium-power work deals with magnifications from about 30 to about 500 diameters, and high-power work extends from 500 diameters upward. The limit of useful magnification is a much disputed question. It is sometimes contended that 1,500 diameters represents about all that is worth while, but the fact that very few pictures are published which exceed 1,500 diameters in magnification would lead to the conclusion that either the limit is from 1,000 to 1,500 or else the art has not been developed to the state where substantial gains result by going higher. This matter will be considered at greater length below.

GENERAL DISCUSSION OF APPARATUS

The reason that photomicrography is grouped under three classifications according to magnification, is because the apparatus used in each case is quite different and because the preparation of the subject and its treatment also differ. In fact for low-power work the microscope often may be dispensed with entirely, the lens being secured directly to the camera; in other cases, the microscope serves only as a convenient support for a lens. In the treatment of most transparent mounts an illuminating device termed a substage condenser is necessary, the microscope then forms a very necessary adjunct to low-power work.

Medium-power work always requires the use of a microscope, and because rigidity in mounting and accuracy in adjustment are very necessary to correct rendering of the image, some sort of a stand is provided on which the microscope and a suitable illuminating train are mounted. Usually this stand takes the form of a narrow wooden or metal table supported by substantial metal legs. The table carries an optical bench which in practice is a metal bar or rail of special and rugged construction upon which the optical parts, the illuminant and the camera are mounted and are capable of adjustment so that they may be aligned optically. The description necessarily, meets generalized conditions. There is, however, a great similarity in the product of different makes of equipment and they all follow the same conventional lines, improvements in one make quite often being met by similar changes on the part of other makers.

There is no very well defined line between medium-power and high-power apparatus so far as the stands are concerned, but when it comes to real precision apparatus the choice in equipment is limited to possibly two or three makes. The difference is to be found in the quality of the optical parts and in the general stability of the assembly. A skilled technician may produce remarkable medium-power results with quite ordinary apparatus but no amount of training and skill can make good in high-power work for the actual shortcomings of an objective. Given a really good objective the skilled operator may use an inferior type of stand and secure very fine results, but he will be working under a considerable handicap and his work will not be consistently good because lack of the right sort of apparatus is apt to introduce variations in illumination, focusing, or adjustments which will prove ruinous to good definition.

Thus far consideration has been given to apparatus capable of yielding a magnified image of some tangible sort of a specimen, but there is an entirely different form of microscopic equipment which reveals the presence of particles beyond reach by all other known means of microscopic vision; reference is made to the ultra-microscope. This instrument is not ordinarily provided with photographic apparatus although with certain classes of work and under favorable conditions it is possible to reproduce the image photographically. Both liquids and solids may be studied by this means but in each case the specimen must be capable of transmitting light.

The Compound Microscope

It is obvious that a complete technical discussion of the instrument and equipment used in photomicrography is not within the scope of this paper nor would it be of interest to many readers. In order to appreciate the possibilities of technical microscopy as an aid in the

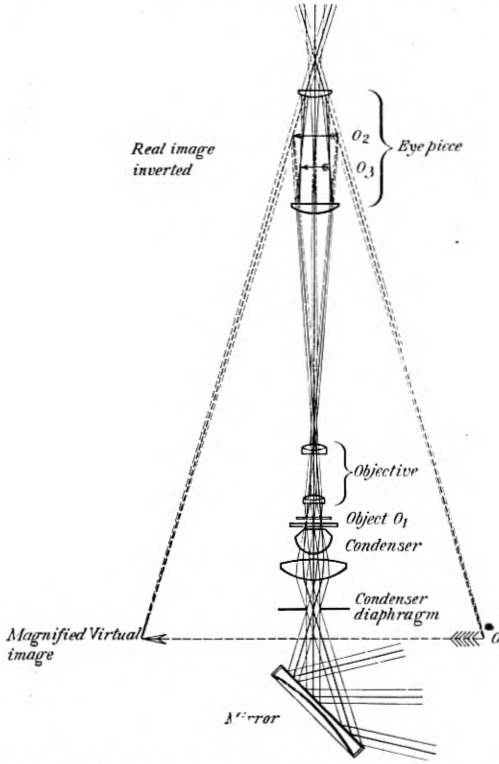


Fig. 1—Optical Diagram of the Compound Microscope.

solution of definite engineering problems relating to telephone apparatus it is necessary, however, to consider more in detail the equipment used.

The optical system of the compound microscope is shown diagrammatically in Fig. 1, and in Figs. 2 and 3 are pictured two modern representative research type microscopes. In the diagram three parallel pencils of light are shown reflected upward into the condenser which by proper focusing is caused to illuminate a transparent object (suitably prepared and mounted as described later) placed in position on the microscope stage. As shown the objective would form an inverted real image of the object O_1 at O_2 but the rays are intercepted by the lower lens of the eyepiece before the real image is formed. The lower eyepiece lens in combination with the upper

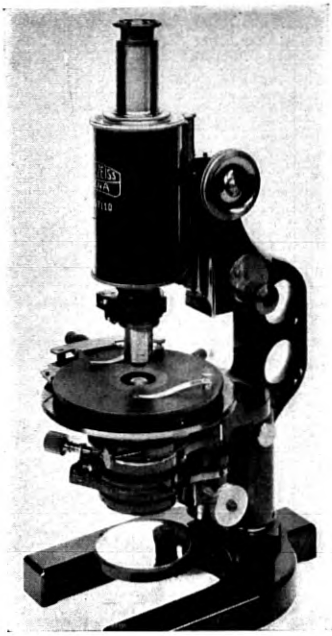


Fig. 2—Research type of microscope by Zeiss. Large barrel for photo-micrography; revolving mechanical stage, and sliding objective changers.

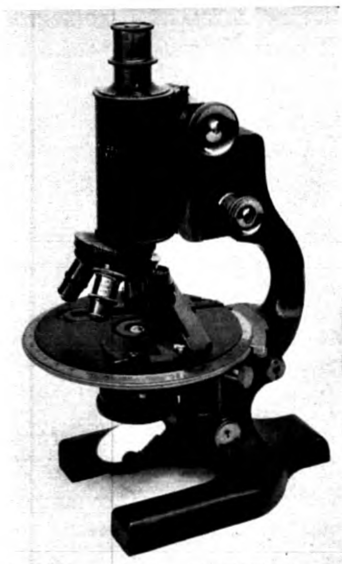


Fig. 3—Research type of microscope by Spencer Lens Co. Large barrel for photo-micrography; a large revolving stage with graduated circle, and a removable mechanical stage.

eyepiece lens forms a magnified virtual image O_4 of the real image O_2. There are two magnifications of the object and the resulting final magnification is the product of the magnifying powers of the objective and the eyepiece.

It should be noted that the objective produces an enlarged image of the object and that the eyepiece further magnifies this image; from this it is evident that if detail is lacking or if the image is not a good likeness of the object, the eyepiece will not make up for the shortcomings of the objective. The objective, then, becomes perhaps the most important part of the whole outfit. No one objective will serve for all purposes because of the limited range throughout which each particular objective is most useful; hence it is necessary to have a whole battery available so that the objective may be selected to suit the requirements of the work.

Objectives are divided into four general classes: achromatic, semi-apochromatic, apochromatic and monochromatic for use with ultraviolet light. These objectives do not consist of single lenses but are composed of two or more lenses very accurately centered and permanently mounted in a metal holder. The component parts of the lens system are chosen with regard to their ability to correct or compensate

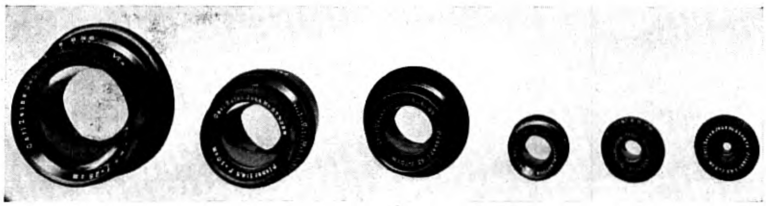


Fig. 4—A battery of low-power lenses. These lenses are used without eyepieces. Each lens is equipped with a diaphragm for stopping the aperture.

for certain errors which are always characteristic of a simple lens. The value of an objective depends on the degree to which these imperfections have been overcome.

The difference in quality between the first three classes of objectives is primarily a matter of correction for chromatic and spherical aberrations. Chromatic aberration is the inability of a lens to focus sharply at the same point the different colors which go to make up the incident light and the inability to bring two rays of incident light of the same color to the same focus is termed spherical aberration.

The achromatic objectives have the chief optical defects corrected in a sufficient degree for the physiologically most effective rays (yellow-green) of the visible spectrum, while in the case of the apochromatic objectives the correction of the image defects extends approximately evenly over the entire range of the visible spectrum from the red to the violet regions.

In the achromatic lenses the fusion of the chromatic rays becomes less and less complete for rays belonging to the extremes of the visible spectrum under the ordinary conditions of illumination with white light, and this imperfection becomes more apparent when highly magnifying eyepieces are used. There are also residual imperfections in the fusion of the rays so that the colors of objects are not rendered with absolute precision in their finer shades. In the apochromatic objectives the fusion of the rays is so perfect that they may be used in conjunction with high-power eyepieces, and because of this perfect fusion the natural colors of the object are rendered with great precision. The semi-apochromatic objectives contain fluorite elements and these objectives occupy a position in quality intermediate between the achromatic and apochromatic types.

Objectives are classified and listed according to their optical characteristics such as primary magnification, numerical aperture and focal length and as to whether they belong to the "dry" or the "im-

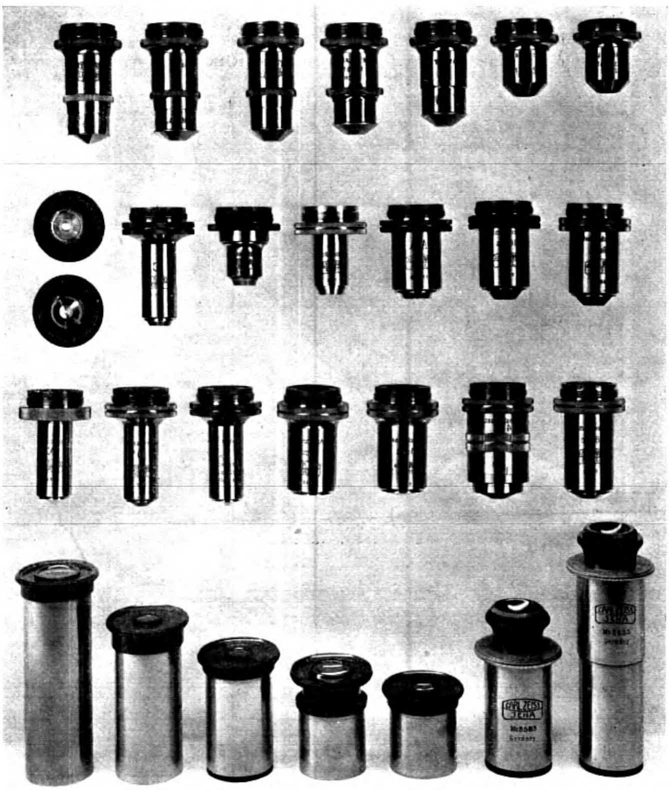


Fig. 5—A battery of medium and high-power objectives and eyepieces.

mersion" series. The term "dry" signifies that the objective when properly used is separated from the specimen by a stratum of air. In the case of the immersion objectives some one fluid for which medium the objective has been computed, such as water, glycerine, cedarwood oil, etc., is used to connect the front lens of the objective with the specimen. The fundamental difference between the dry and the immersion objectives is one of resolution, where by resolution is meant the ability to see separate and distinct lines as individual units when these lines are spaced very close together. Resolving power or the number of lines per inch resolved is expressed numerically by the equation

$$N = \frac{2\ N.A.}{\lambda},$$

in which N is the number of lines per inch, $N.A.$ is the numerical aperture (defined below) and λ is the wave-length in inches. An objective of high resolving power when correctly used will resolve

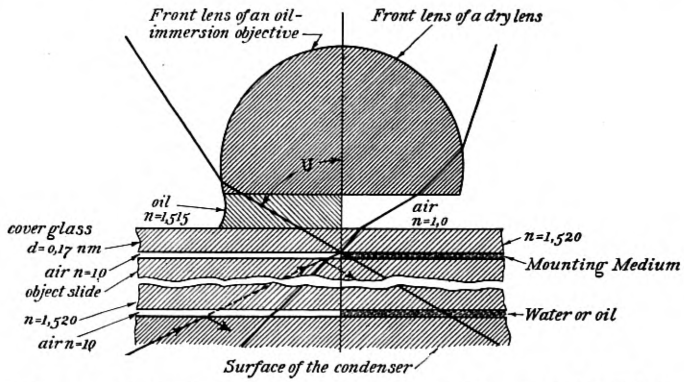


Fig. 6—Diagram illustrating numerical aperture and the superior light gathering powers of an oil immersion objective.

lines spaced 100,000 to the inch, whereas an objective of inferior resolving power under the same condition will not be able to distinguish the lines as distinct units.

As will be seen from Fig. 6, an immersion lens has greater light gathering power than a dry lens of corresponding focal length. This light gathering power is expressed as numerical aperture which term in reality supplies a measure of all of the essential qualities of the objective. The magnitude of the numerical aperture is expressed by the equation

$$N.A. = n \sin U,$$

n being the refractive index of the medium contained between the cover-glass and the front lens of the objective, and U the semi-aperture angle of the system.

For a given magnification and under comparable conditions the resolving power is directly proportional to the numerical aperture. The brightness of the image is proportional to the square of the numerical aperture. As the numerical aperture increases the depth of penetration (*i.e.*, the power of the objective to resolve detail simultaneously at different depths or distances from the objective), and the flatness of the field both decrease, but usually when high resolution is desired flatness of field and penetration are not of great concern. The value of the numerical aperture varies from about 0.10 in the very low-power achromatic objectives to 1.40 for the oil immersion apochromats.

The eyepieces for use with the achromatic objectives are generally of the Huygens type but those for use with apochromats are termed compensating because of certain corrective measures which they apply to the behavior of this type objective. High-power achromatic objectives and the semi-apochromats may also be used to advantage with the compensating eyepieces. The magnifying power of eyepieces ranges from about 4 times to 20 times although another class termed orthoscopic eyepieces may be procured with a magnifying power of 28. These latter eyepieces are generally used with low-power objectives only. A special type of eyepiece known as a projection eyepiece of low magnifying power is used for certain classes of work when photographing with a long bellows extension. These eyepieces have correction collars which must be set to correspond with the bellows extension used.

Illumination

The color of the light used and the illumination of the specimen play a most important part in photomicrography and the behavior of the finest objective will appear very ordinary unless critical illumination of the specimen is attained. The illuminant is usually some form of arc lamp or metal filament, gas-filled lamp. Both types have their advantages and while many statements may be found derogatory to the use of arc lamps, as a real source of light, the author has found that a smoothly operating automatic arc lamp equipped with suitable carbons is capable of yielding results of the highest order. What is needed especially for medium and high-power work is a point source of light (or approximately so) of great brilliancy; capable of being smoothly and uniformly controlled so that the luminous end of the positive carbon will not fluctuate backward and forward within wide limits. Most automatic arc lamps are designed for a certain direct current value, usually about five amperes and unless the current rating is closely adhered to in practice the operation is apt to be irregular. Sputtering and irregular feeding of the carbons are due to lack of proper adjustment of apparatus or of current and voltage or to the use of an unsatisfactory grade of carbons. It may be of interest to know that the type of automatic arc lamps used in the Bell System Laboratory are so steady and uniform in their operation that they occasion no concern except for the usual maintenance. Since very small diameter carbons must be used to approximate the point source of light condition, these lamps will operate continuously with one set of carbons for about thirty minutes

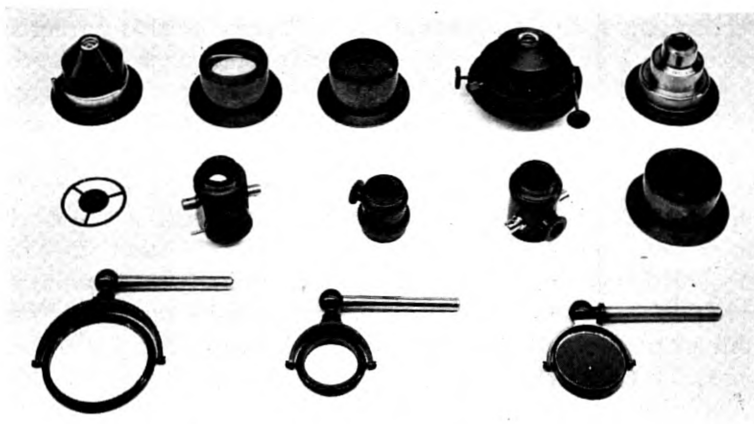


Fig. 7—Condensers and illuminators used in microscopy

only. Frequently exposures are made at high powers (6,000 to 9,000 diameters) lasting from 45 seconds to 3 minutes, during which time the carbons may feed several times and no ill effects result, so perfect is the operation.

Critical illumination is nothing more than bringing the rays of light from the source of illumination to a state of proper focus and optical alignment so that the surface of the specimen under examination will be uniformly and brilliantly illuminated. This matter of securing uniform illumination is by no means the simple operation that the designation implies since usually an optical train consisting of the light source, condenser, diaphragms and an object illuminating device of some sort must all be brought into exact optical alignment with the optical system of the microscope.

For very low-power work or for gross photography of specimens, a gas-filled, metal filament lamp with a suitable condenser and mounted on a portable pedestal which may be adjusted to all angles is very useful. In this case the optical train is dispensed with and the light thrown at the proper angle on the specimen to be photographed.

Great brilliancy is not required for this work but rather a diffused light, obtained by means of interposing a ground glass screen in the illuminating beam.

The object in photomicrography is to record as clearly and as faithfully as possible the structural characteristics of the specimen. This is accomplished by a rendering of contrast between the structural elements of the specimen and by intensifying or diminishing this contrast to suit the particular characteristics which are to be reproduced to best advantage. This control of contrast is obtained by control of the color of the light used for illumination.

A spectroscopic analysis of the light of the arc shows a continuous spectrum consisting of three dominant color portions, blue-violet, green and red which pass by gradation to each other; the blue-violet passes by blue and blue-green to green, and the green by yellow and orange to red.

If an object absorbs some constituent of the white light falling on it then the reflected light will be deficient in this color and as a result the eye will experience the sensation of color.

The effect on the color of the residual light by blocking out a narrow band at different positions in the spectrum is shown in Fig. 7a.

Fig. 8—Diagram representing the spectrum of arc light divided roughly into three dominant bands.

A simple diagrammatic representation of the visible spectrum is shown in Fig. 8, in which the tri-color division is broadly made as follows:

Blue-Violet	4,000 to 5,000 A.U.
Green	5,000 to 6,000 "
Red	6,000 to 7,000 "

An object which appears red to the eye when illuminated by white light is absorbing the blue-violet and the green light, and the bulk of what it reflects or transmits is red. Similarly, an object appears green because it is reflecting or transmitting the green constituents of the spectrum and absorbing the red and the blue-violet rays. These are simple cases assuming sharp absorptions and ideal conditions, but in the practice of the art of photomicrography we are dealing with gradation in color and oftentimes the structural characteristics of the specimen show little contrast, either within the specimen itself, or between the specimen and the background. Therefore, to reproduce an object faithfully or to accentuate faintly revealed characteristics, careful consideration must be given to the color of the light used when photographing the specimen. For the purpose of separating white light into well defined bands, light filters are used and their function is to filter out rays or bands of rays of certain given wave-lengths. These filters consist of colored gelatine films mounted between flat pieces of glass or of liquids appropriately colored and contained in rectangular vessels of glass with flat and

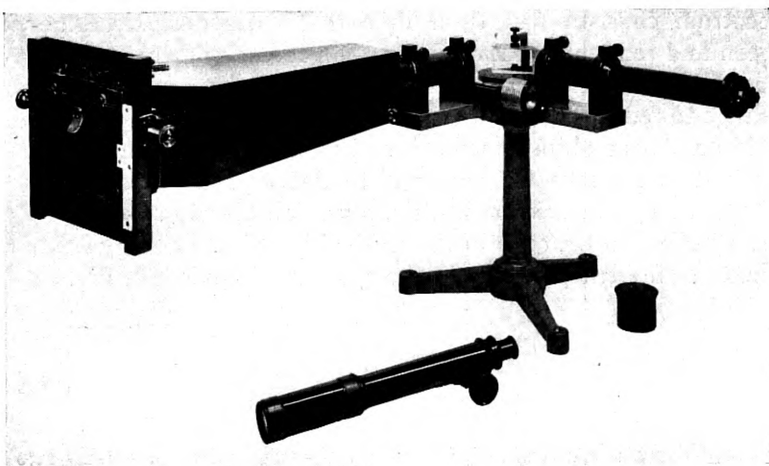


Fig. 9—Hilger wave-length spectrometer. The camera is interchangeably mounted with a reading telescope.

parallel side walls. The "Wratten M" series of gelatine and glass filters is probably the best known and most widely used. The selection of a light filter for a given specimen is usually by experimental methods. Successive filters are inserted in the illuminating beam and the resulting image studied for rendering and definition. However, two simple rules apply generally; if a color is to be rendered as black as possible, then it must be photographed by light of wave-lengths within the absorption band of the specimen; when contrast is desired within the specimen itself, the object should be photographed by light of a wave-length which it transmits. The first rule is of use when it is desired to secure contrast between the object and the background; and the second for better rendering of detail within the object.

Spectrometers are available for determining the characteristics of filters; for determining the transmission spectrum of a micro-

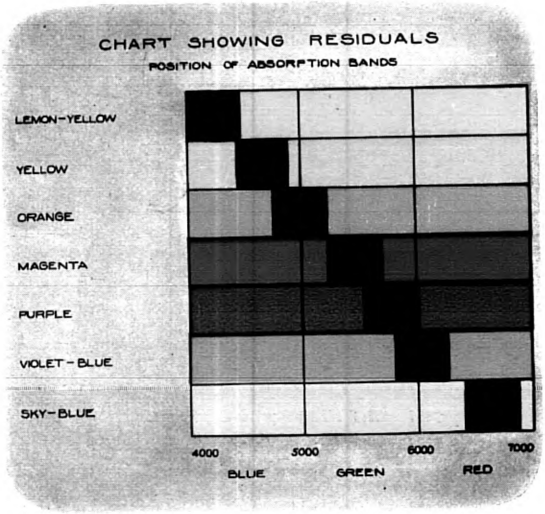


Fig. 7a—The effect on the residual color of arc light by blocking out narrow bands at different positions in the spectrum.

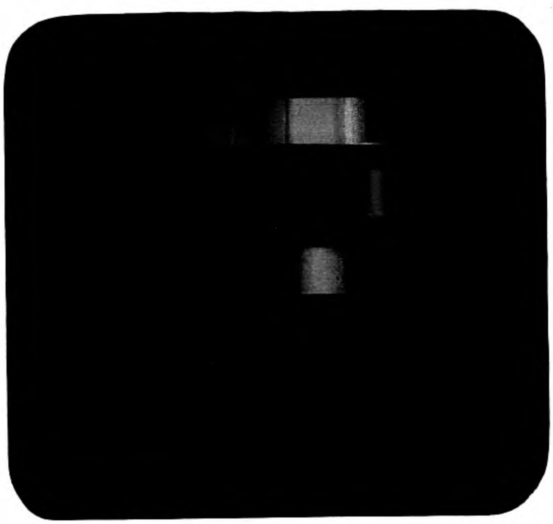


Fig. 10—Direct color photography of the spectrum of arc light and with the Wratten A, B, C and D filters respectively interposed in the beam.

scopic mount; and for studying the effect of dyes or stains on certain types of transparent mounts. For the purpose of filter studies the Hilger wave-length spectrometer constitutes a very useful accessory. This instrument is illustrated in Fig. 9, and in Fig. 10 is shown by direct color photography the residual light from an arc lamp after passing through various filters. The spectrometer is adapted for either direct vision work or photography, a camera and telescope being interchangeable. Instruments for observation with spectroscopically decomposed light constitute what are known as spectroscopic eyepieces and are very useful for certain classes of work, since they replace the usual microscopic eyepiece and may be used with any objective. Precision instruments of this type are capable of measuring the transmission or absorption spectrum of very minute

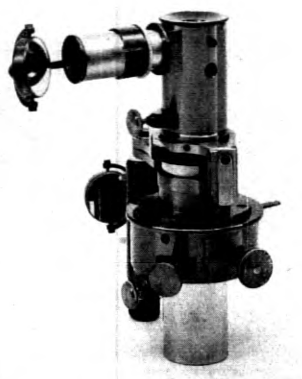


Fig. 11—A spectroscopic eyepiece by Zeiss. This instrument replaces the usual eyepiece of the microscope when it is desired to make observations with spectroscopically decomposed light. It yields an image of the transmission spectrum of the object with a superimposed Angstrom scale and if desired the transmission spectrum of the staining reagent may also be brought into the field of vision. The staining reagents are placed in glass vials.

bodies such for example, as a single blood corpuscle, which state of perfection is said to be attained by the Zeiss instrument, illustrated in Fig. 11.

The wave-length of the light used in photomicrography also has other useful functions to perform and for some classes of work these take precedence. Mention has been made of the correction of objectives for aberrations which are inherent in the simple lens. When an objective, not fully corrected, is used for photomicrography at the higher magnifications, color distortions assert themselves and result in faulty performance of the objective unless filters are used to exclude light of wave-lengths other than that for which the objective has been computed.

In high-power photomicrography of metallurgical specimens, the purpose, of course, is to attain the maximum of resolution and here the wave-length of the light used plays an important part. As mentioned above the resolving power of an objective may be increased by decreasing the wave-length of the light used. Assuming that a Wratten "F" filter is used whose transmission band is from 6,100 A.U. to the red end of the spectrum, then an objective of 1.4 N.A. should resolve about 109,000 lines per inch. If a "C" filter is used whose spectral transmission is from 4,000 A.U. to 5,100 A.U., the same objective should resolve about 158,000 lines per inch. In other words, by using the shorter wave-length light,[1] it is possible to effect a theoretical improvement of about 45% in the resolution. In practice, these theoretical values are not fully obtained because of other complications entering into the problem.

Polarized Light

Polarized light is oftentimes a very useful aid in the study of transparent objects. By combination with suitable selenite plates color combinations are developed in the specimen and between the specimen and the background which facilitate identification of substances, comparison of known and unknown substances, and the study of their structure. In the field of crystal studies, polarized light is indispensable and it furnishes evidence of a very substantial nature in the field of micro-chemistry. The problem has been presented on occasions to identify the nature of some substance, resulting from the corrosion of some small telephone part. The evidence in these cases could easily be placed on the head of a pin but by the use of polarized light in conjunction with micro-chemical methods, it has been possible to form some sort of a qualitative estimate of the nature of the substance. Polarized light is obtained by means of a nicol prism contained in a suitable mount which is clamped in a ring beneath the sub-stage condenser. The illuminating beam from the microscope mirror is thus polarized before it reaches the condenser. A second nicol prism called the analyser is either contained within a special eyepiece or the analyser takes the form of a mount which may be placed above the usual eyepiece. Both polarizer and analyser are

[1] Regarding the use of ultra-violet light see High-Power Photomicrography of Metallurgical Specimens, F. F. Lucas, Trans. Am. Soc. for Steel Treating; Vol. IV, p. 611, 1923.

mounted so that they may be revolved. Extinction angles are read from a suitably graduated circle usually forming a part of the analyzing eyepiece.

Preparation of Specimens

Specimens, to be investigated or studied by microscopic methods must have a preparatory treatment in all cases except, perhaps, for very low-power work. Many samples require the preparation of transparent sections: that is, a specimen of the object a few thousandths of a millimeter in thickness so that it is transparent or at least translucent; studies of woods, porcelains, papers, fibers, tissues, insulating compounds, etc., are usually made with transparent sections.

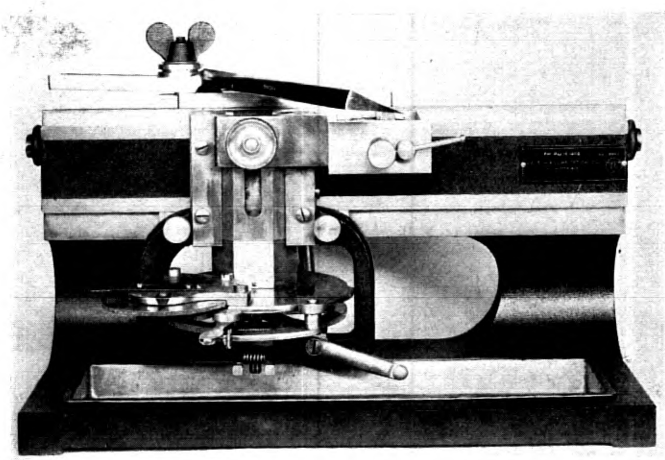


Fig. 12—A sliding microtome for cutting microscopic sections. The work is held in a clamp and a very heavy section razor, flat on one side and hollow ground on the other is operated backward and forward on a slide rails. The return movement of the razor operates the elevating mechanism to which the work is attached so that the latter may be raised to cutting position by predetermined increments.

Hard specimens such as porcelain are prepared by grinding, softer materials such as wood sections are first prepared by suitable softening processes and then are cut in an instrument called a microtome, a form of which is shown in Fig. 12.

Delicate structures require strengthening before they can be cut; these are embedded in paraffin, celloidin or glycerine gum. For successful results gradual and thorough impregnation of the parts is required and this operation may take several weeks. After the sections are cut, they must be further prepared by being stained, dehydrated and cleared after which they are finally mounted in Canada Balsam or similar mounting medium between a glass microscope slide and a cover glass. Mounts of this kind are permanent, but when it is not desired to retain the mounted specimen for record or future examination, temporary mounts are often made in which the mounting medium is some liquid such as water or glycerine, or

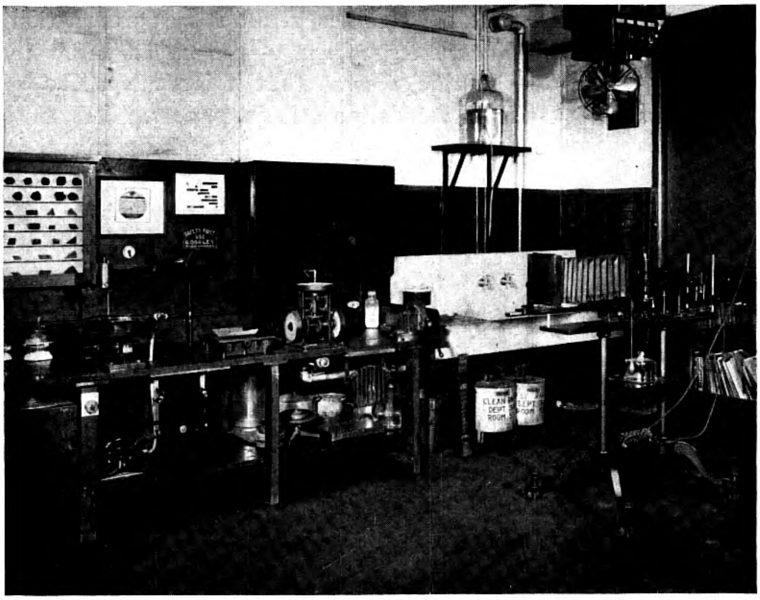
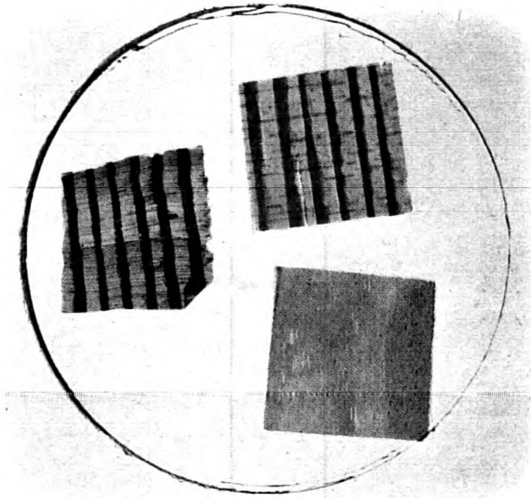


Fig. 14—Equipment for the preparation and preliminary examination of opaque specimens.

in some cases, may be the staining medium itself. An enlarged view of a permanently mounted transverse radial and tangential sections of Douglas Fir wood is illustrated in Fig. 13.

The preparation of metallurgical specimens is accomplished by different methods and if a specimen is to be examined at extremely high powers, the utmost in skill and refinement of methods is necessary. The usual method of procedure is first to file a flat surface on the specimen, after which the surface is gradually brought to a semi-polished condition by rubbing the specimen on a sheet of French emery paper, placed on a plane surface. A coarse grade of paper is first employed and by gradual steps, finer and finer grade papers are used, the rubbing on each successive paper being in a direction at right angles to the preceding paper and continued until the scratches



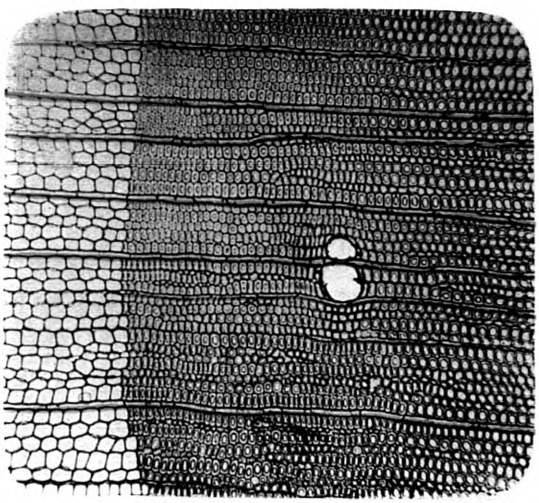


Fig. 13—An enlarged view of a specimen prepared for microscopic examination. The cover glass shown by the circle measures ¾ inch in diameter. The mounts are transverse, radial and tangential sections of a wood specimen and were stained to make their structure visible under the microscope. The appearance of a transverse section of Douglas Fir at 100 diameters is shown in the lower illustration.

of the preceding operation have all been removed and finer ones established in the new direction. This is continued to the 000 paper, after which the specimen is further polished on a polishing machine having a broadcloth covered lap capable of being revolved at varying speeds to about 1,200 rpm. This lap is kept moistened with water and fine alundum is used as the abrasive. This operation gives a

Fig. 15—General view of the Laboratory for Technical Microscopy.

semi-polish and when properly carried out, leaves the specimen with numerous very fine scratches. The final operation is carried out on another lap covered with very fine broadcloth and with an exceedingly fine abrasive such as the finest jeweler's rouge or magnesium oxide. For high-power work magnesium oxide is the only polishing medium which has been found to yield a satisfactory surface. The technique for the development of surfaces at high powers has been worked out in our laboratory so that it is now possible to study metal structures with great clearness at high powers. Equipment for grinding and polishing specimens is shown in Fig. 14.

Metals, after polishing, as a rule, do not show their structural characteristics, but must be treated in some way to etch the polished surface. This etching operation is a simple matter for low-power work, but as the magnification is carried higher and higher, the problem becomes increasingly difficult.

Bell System Photomicrographic Laboratory

A general view of the Laboratory situated on the fifth floor of the building at 463 West Street, New York City, is shown in Fig. 15. Some of the equipment is more fully illustrated in detail views.

It consists of two metallurgical equipments, one of which is the large Zeiss metallographic outfit shown in the foreground of Fig. 15. This equipment is of precision quality and is used for all classes of

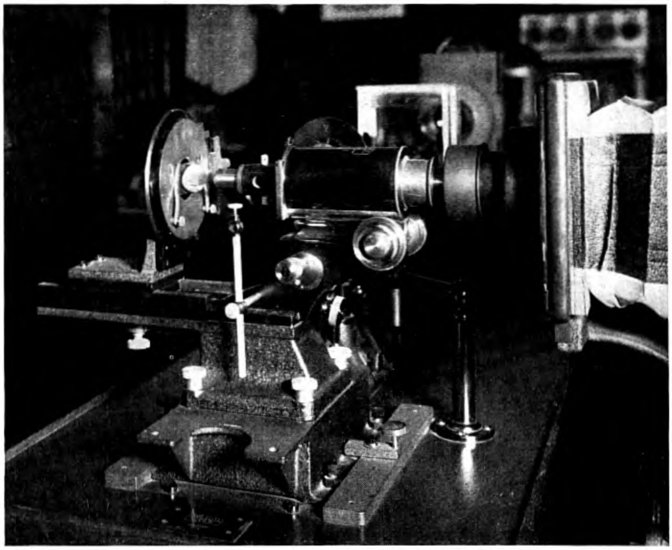


Fig. 16—The Martens stand of the large metallographic outfit. The vertical illuminator is shown between the barrel of the microscope and the objective.

work involving opaque specimens. The optical parts consist of a full complement of Zeiss apochromatic objectives and compensating eyepieces for medium and high-power work. For low-power work a full set of Zeiss micro-planar lenses and a Tessar lens are used. The maximum bellows extension of the camera is 155 centimeters and the plate holders are designed for 24 x 30 centimeter plates and all smaller sizes by employing suitable kits. The illuminating train consists of an automatic arc lamp, a condensing system, and cooling cells, mounted on an optical bench and capable of adjustment to meet the conditions of the work.

Illuminators of conventional types, for vertical and oblique light may be assembled on the Martens type stand. This stand is a departure from the construction of the usual form of microscope stand. It is much more rugged and is arranged for use in a horizontal position only. In precision work a stand must be stable and substantial and the construction throughout has been arranged with this thought in mind. The microscope is equipped with a movable stage for rough focusing and this is fitted with a revolving mechanical stage so that the specimen while under examination may be moved about at will for the purpose of study or exploration. To facilitate focusing,

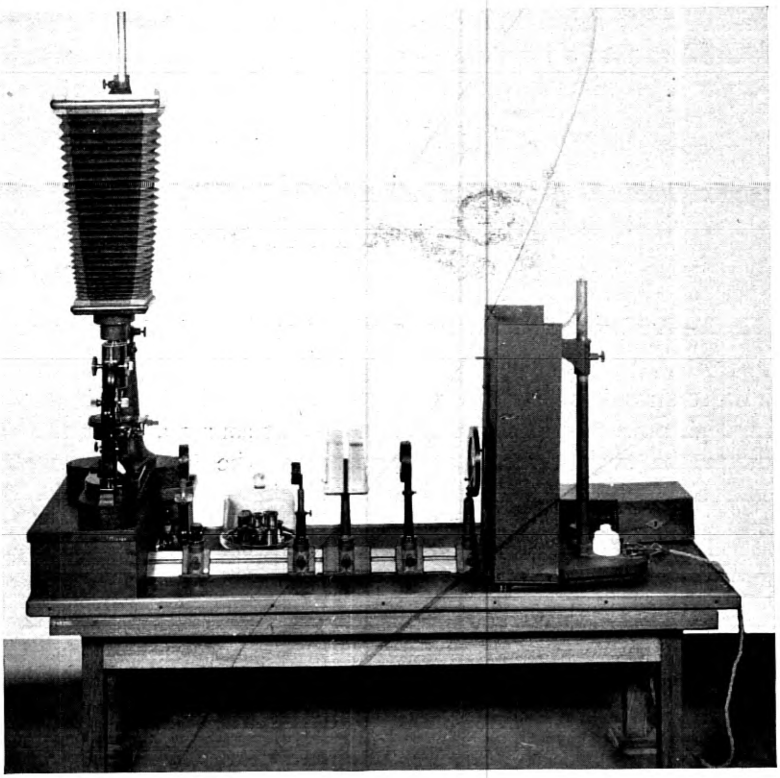


Fig. 17.—A vertical photomicrographic camera for transparent specimens.

gear is provided so that the operator may sit at the ground glass screen and by means of a wooden handle, focus the microscope. A ground glass screen for viewing and for rough focusing and a clear glass screen for fine focusing with a magnifier are provided to be interchangeably mounted with plate holders on the camera back.

A second metallurgical outfit of Bausch and Lomb manufacture shown in Fig. 14, is used for preliminary examination of specimens while in the course of preparation and for photographing some metallurgical specimens at medium powers. This outfit is also arranged for photomicrography and has a 5 x 7 camera of rather short bellows extension. The objective equipment is of the achromatic type.

For transparent work a Zeiss vertical camera outfit, Fig. 17, equipped with the conventional Zeiss research type microscope is used. The camera has a bellows extension of 80 centimeters and uses 5 x 7 or smaller plates. It is fitted with ground and clear glass focusing screens similar to the large Zeiss metallurgical outfit. The illuminant is a 500 watt metal filament nitrogen filled bulb with the filament mounted so that a large circular area of illumination is presented, or if desired, the filament assembly may be turned sideways and a single

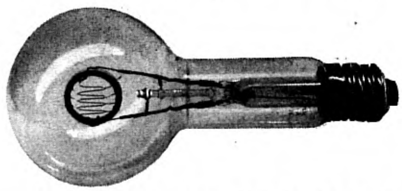


Fig. 18—A 500 watt metal filament, gas-filled lamp for use in photomicrography.

filament strand is thus presented to the optical train. In medium and high powers, this approximates a point source and for the lower powers the large circular arrangement of the filament provides a relatively large area of illumination which is quite desirable. The lamp is illustrated in Fig. 18. The illuminating train consists of condensing and cooling units adjustably mounted on a substantial optical bench as in the case of the metallographic outfit. The objectives consist of a full set of apochromats and also several achromats of low power. The micro-planars are also used with this equipment.

The Ultra-Microscope

The ultra-microscope is an instrument for revealing the presence of very minute bodies present as colloids in transparent solids or liquids. The presence of these particles is made apparent by the light rays which they intercept and diffract upward into the microscope objective. It is a matter of common observation that dust particles are seen in an intense beam of light such as sunlight but otherwise their presence remains concealed. This principle of illumination is made use of in the ultra-microscope as described below and accordingly differs considerably from the conventional arrangement of compound microscope and illuminant.

The appearance of ultra-microscopic particles in fluids and transparent solids as seen by means of the ultra-microscope is, without a doubt, one of the most fascinating and spectacular demonstrations within the scope of technical microscopy. A beaker containing water with a drop or two of glue or soap, or containing benzol with a few drops of a rubber solution stirred into it, or even some rather dirty looking oil which has seen service in some machine, do not constitute

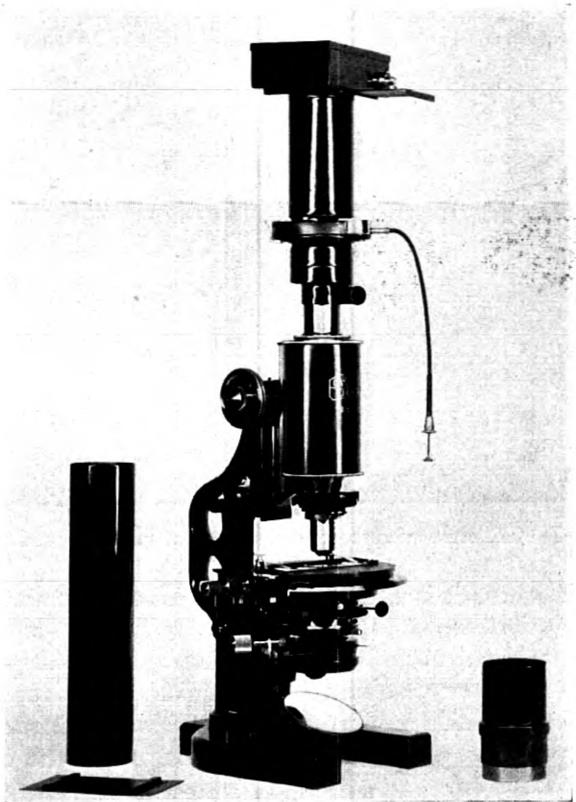


Fig. 19—A small photomicrographic camera developed by the writer and used extensively in the laboratory for photographing on film, or on plates. It is used when a large number of small specimens are to be reproduced or when a large field is unnecessary.

interesting exhibits as viewed in the beakers, but placed in suitable cells for ultra-microscopic examination, these liquids come to life and display the colloidal particles coming into vision as tiny illuminated particles, only to burst into rings of light and pass away into the dark background. The constant irregular motion is the Brownian movement and the smaller the particle the more lively it moves. Conglomerate masses of particles merely float through the field of vision and, compared with the individual particles, appear exceedingly sluggish.

Fig. 20 gives a general view of the Zeiss ultra-microscope as originally devised by Siedentopf and Zsigmondy. The equipment has been superseded to some extent by the later Siedentopf cardioid ultra-microscope. The latter is a very powerful light-concentrating device

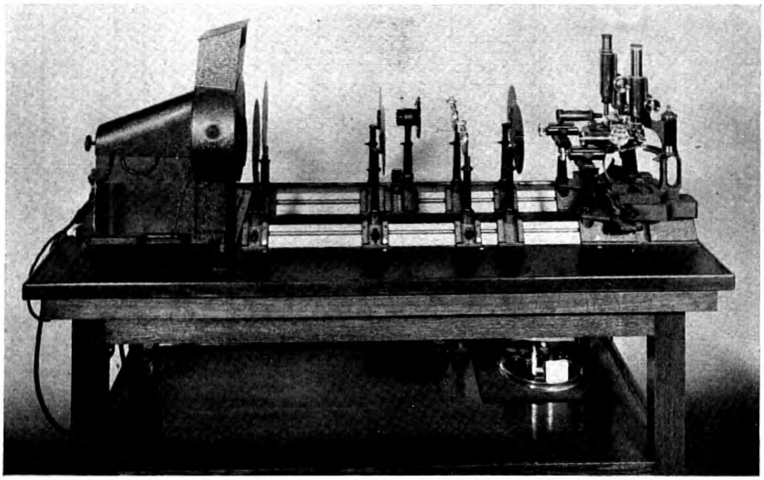


Fig. 20—The Slit ultra-microscope for transparent solid or liquid specimens.

and for this reason it is primarily adapted for the examination of fine colloidal solutions and dilute precipitates as well as for the observation of micro-chemical and photo reactions. For transparent solids and for the precursory examination of liquids and for rapidly passing in review several fluids in succession, the original arrangement retains marked advantages. The cardioid ultra-microscope will be described more fully later on.

Fig. 21 shows diagrammatically the path of the rays within the preparation in the presence of ultra-microscopic particles and will serve to make clearer what is to follow. In the original form of ultra-microscope (Fig. 20) the horizontal incident rays which go to furnish the illumination do not enter the microscope, the latter being set up vertically and hence the background appears dark. The only rays to enter the objective of the viewing microscope are the diffracted rays which come within the aperture of the objective.

At one end of the base board is an automatic arc lamp mounted on slide rails so that it may be brought in line with either of two illuminating trains mounted on optical benches.

One of these illuminating trains functions with a microscope which has mounted on its objective a clamping device for holding Biltz cells in which the liquids are placed for examination. The other train serves another microscope on the stage of which is mounted a special object stage capable of being raised and lowered and provided

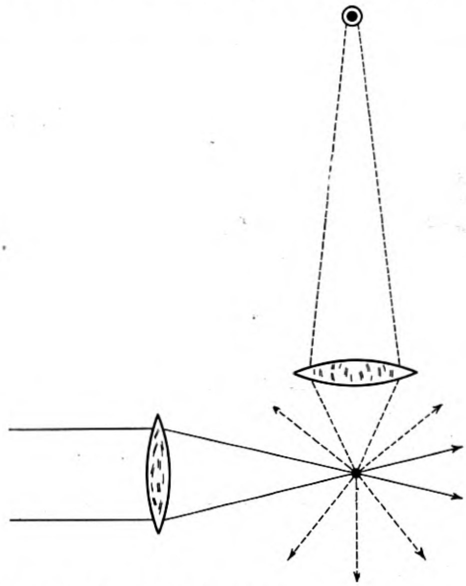


Fig. 21—Illustrating diffraction of light impinging upon an ultra-microscopic particle. Illuminating rays represented by solid lines and diffracted rays by dotted lines.

with a plate at the top to receive the specimen to be examined. In this case the specimen, if a hard solid, has been previously prepared to have two ground and polished surfaces in planes at right angles to each other and is mounted so that one faces the illuminating train and the other the objective of the viewing microscope. Plastic substances or certain liquids not suited to the use of the Biltz cells are placed in a special glass cell having a deep cylindrical recess faced with a quartz window toward the illuminating train. Various cells for ultra-microscopic examinations are shown in Fig. 22.

Placed next to the arc lamp is a fixed diaphragm and then a small projection lens which is corrected chromatically and spherically and brings the image of the positive carbon of the arc lamp to a focus on the adjustable slit. The slit is provided with a drum bearing a scale. The divisions of the scale embrace 50 parts and a complete revolution of the drum opens the slit ½ mm. so that each division of the scale advances the slit 1/100 mm. The slit is fitted with two jaws at right angles to the principal slit, one being adjustable by a milled screw head. The function of these jaws is to limit the length of the slit. The slit head may be given a quarter turn so that it may be set horizontally or vertically, which is necessary in order to calibrate the instrument as explained later. A projection lens next in order toward the microscope projects the image of the slit into the image plane of a horizontally mounted objective which is mounted on a stand with cross

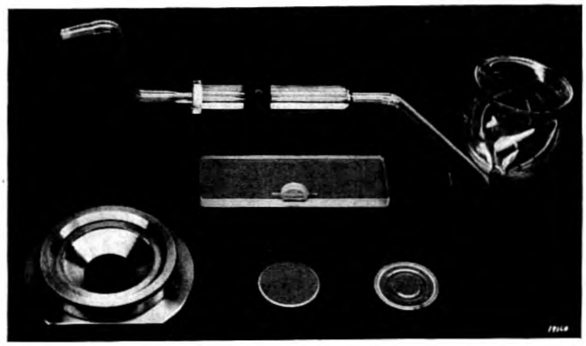


Fig. 22—Cells used for the examination of fluids with the ultra-microscope.

slides so that the objective which serves as an illuminator may be moved horizontally in two directions, at right angles to each other. The movement of the cross slides is controlled by screw adjustment but for coarse adjustment in the direction of the illuminating train a sliding sleeve adjustment is made. By this means the illuminating objective can be centered with respect to the observing microscope objective. In the correct position the front lens of the illuminating objective is about 1 mm. from the mount of the observing objective.

The Biltz cell has a rectangular cross section which permits of accurately adjusting the cell in position. A thistle funnel at one end is for the reception of the liquids; the other end is provided with a piece of rubber tubing which has a pinchcock to prevent the escape of the fluid. The rectangular section of the cell has two quartz windows, one of which normally faces the illuminating objective and the other the observing objective. The cell is attached to the observing objective by means of the clamp mentioned and the cell is focused in the proper position in the beam of light by racking the microscope draw-tube upward or downward in the usual manner by the coarse and fine adjustments. The observing objective is a special water immersion objective which makes contact with the upper window of the Biltz cell through the medium of a drop of distilled water.

Quantitative investigations are made by counting the visible particles in a given volume of the fluid and the manner in which so novel an investigation can be accomplished by optical means should prove of general interest. One method consists in the use of the eyepiece micrometer which is substituted for the ordinary eyepiece of the observing microscope. The eyepiece micrometer is ruled into squares and the dimensions of these are found by calibration with a stage micrometer. The depth of the stratum is measured by turning the slit head through a right angle and thus a solid is blocked out in the path of the light rays, whose length and breadth are defined by the rectangular area of the micrometer eyepiece and whose depth is

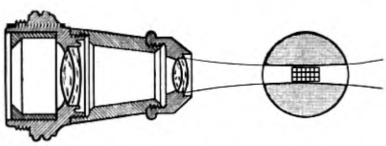


Fig. 23—Illustrating the adaptation of micrometry to the ultra-microscope for the purpose of counting particles per unit volume.

that of the light beam and may be read from the known dimensions of the eyepiece micrometer. Fig. 23 shows the cross ruling of the eyepiece and the pencil of light which traverses the field. The length of the side of each square as seen through the water immersion objective with a tube length of 160 mm. has an approximate value of 9 μ as referred to the object, which value is sufficiently accurate for ordinary measurements. Where more exact measurements are required, the ruling is calibrated for the eyepiece and objective by means of a stage micrometer in the manner to be described under the subject of micrometry.

For studying the behavior of particles in polarized light the eyepiece is fitted with an analyser. In a measure, as the particles decrease in size they become more strongly polarized in degree towards the plane which passes through the axis of the illuminating and diffracted rays, i.e., the principal plane of diffraction. The analyser also serves to distinguish unpolarized from polarized light.

The apparatus for examination of solids is identical in so far as the illuminating train is concerned with the apparatus for liquids just described. It differs only in the character of the specimen or of the cell used and, while designed primarily for transparent solids, it may be used with a suitable cell for liquids also. When liquids are being examined, there is no need for searching the specimen since the particles are in constant motion, but when solids or semi-solids are being examined it may be desirable to do so. The mechanical stage of the microscope on which is mounted the adjustable specimen stage allows any layer in the specimen to be brought into accurate focus and hence various strata of the specimen can be examined one

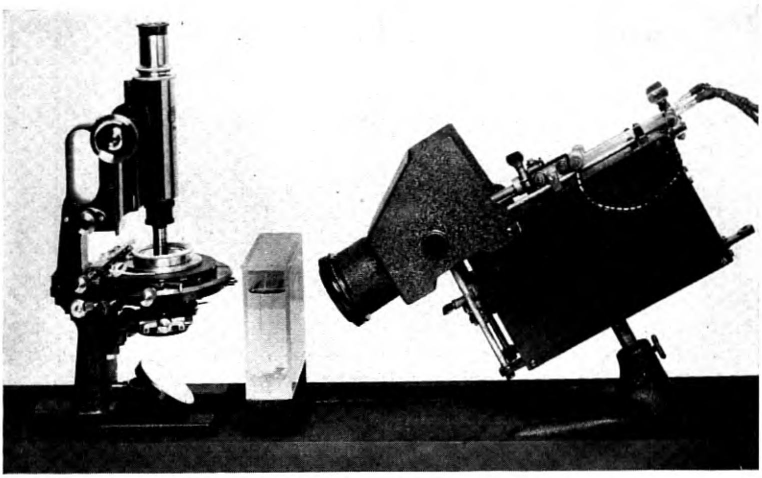


Fig. 24—Cardioid ultra-microscope.

after the other. As previously stated, the specimen must be provided with two polished surfaces at right angles to each other to correspond to the quartz windows of the Biltz cell.

Since the observation of ultra-microscopic particles in polarized light supplies useful information respecting their form and color, a polarizer is provided with a hinged stand so that it may be swung out of the optical train. The analyser, as previously mentioned, is fitted over the eyepiece of the microscope.

The cardioid ultra-microscope illustrated in Fig. 24 differs only in two important features from the ordinary form of microscope. The illumination of the fluid under examination is obtained by a dark-ground condenser mounted in the sub-stage condenser collar and to which Zeiss has given the distinctive name "cardioid condenser." A diagram of the condenser and the paths taken by the rays is illustrated in Fig. 25. Since the aperture of the rays brought to a focus by the condenser exceeds 1.0, it follows that no light can emerge from the condenser if there be a stratum of air above the condenser. It is therefore necessary to connect the object slide or cell chamber and the top of the condenser by a stratum of immersion fluid free from air bubbles. Cedar oil or pure water is used for this purpose. The chamber for the cardioid condenser is illustrated in Fig. 22. The chamber is made of quartz glass and consists of a circular disc having on one side a circular groove and an optically plain central portion within the groove about 2 μ below the plane outside the groove. A drop of the fluid to be examined is placed on this depressed central portion and a cover glass of quartz placed over it. The excess fluid is expelled to the annular groove and a stratum about 2 μ in thickness is retained in the central portion of the chamber for microscopic examination. The cell is assembled in the metal mount which has a clamping ring and a recessed member

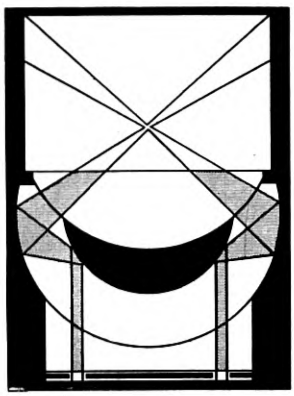


Fig. 25—Diagram of the rays in a cardioid condenser.

to receive it. The very brilliant illumination resulting from the cardioid condenser would cause glass to fluoresce and for this reason a quartz cell is used. Moreover, glass is more liable to be affected by corroding agents than is quartz.

The utmost care must be taken to prepare the cell chamber. This includes washing with alcohol and water; dipping in boiling sulphuric and chromic acid solution; washing in tap water; rinsing in distilled water and then in redistilled alcohol; drying in a hot air current and finally cooling under a bell jar; all of which is necessary to insure absolute cleanliness.

An automatic arc lamp is used as a source of illumination and the image of the crater is projected by a projection lens onto the mirror of the microscope from which the rays are reflected upward into the cardioid condenser.

The objective used with the cardioid condenser is a special apochromat 3 mm. 0.85 N.A. glycerine immersion lens which constitutes a homogeneous immersion lens for cover glasses of fused quartz. This type of objective is necessary because the success of the observation is then largely independent of impurities and slight blemishes on the upper surface of the cover glass, moreover, the lens confers a greater immunity from the effects of varying cover-glass thickness and the immersion fluid precludes the entrance of dust which would gradually cloud the image.

Slit ultra-microscopes are not arranged for photography because in the case of liquids the particles are in a rapid state of motion and the illumination is insufficient. Since in transparent solids the particles are stationary, the image seen in the slit ultra-microscope may be reproduced by making a lengthy exposure. With a small photomicrographic camera developed by the writer the image seen in the slit ultra-microscope for solids has been reproduced and, by instantaneous photography, the moving particles in liquids as seen in the cardioid instrument. Except for purposes of evidence or record, there is little to be gained by photographing with the ultra-microscope.

Dark-Ground Illumination

The dark-ground illuminator constitutes another aid to microscopic investigation. This, in reality, is a sort of ultra-microscope, since the objects are viewed by diffracted light much in the same way as in the cardioid type of equipment. This method of illumination is accomplished by stopping out the axial rays and allowing those of greater aperture to strike the specimen at an angle. The usual form of condenser may be made to yield dark-ground illumination by the simple expedient of inserting a central stop in the path of the light rays just below the sub-stage condenser in a ring provided for such purpose. Better results are attained by use of dark-ground illuminators which are special condensers designed with this object in mind. Dark-ground illumination furnishes valuable means for bringing into view objects which are smaller than about 1 μ. Examples of such objects are furnished by fibers, fine crystalline needles, fissures, edges, rods, bacteria, etc. Under dark-ground illumination methods, these objects are easily seen and studied, whereas with transmitted light, they can be seen with difficulty unless rendered distinguishable by staining. Certain bodies with laminar markings are also suitable subjects for dark-ground studies and in this case the markings are distinguishable more by reason of dissimilarities in refraction than by differences in coloring.

Micrometry

Micrometry plays an important part in technical microscopy because the dimensions of micro-constituents in a specimen are helpful for purposes of identification or for forecasting physical properties. In metallography the measurement of grain size is assuming importance for certain alloys and in some cases, specifications are so drafted as to define this characteristic.

For measuring objects under the microscope, the eyepiece contains a glass disc on which fine divisions have been ruled. In some cases, these rulings take the form of a cross-section composed of small squares or rectangles. The reading of each division of the eyepiece micrometer is calibrated for each objective by comparison with a standardized stage micrometer. These stage micrometers are glass microscope slides on which known units of length have been accurately ruled, such as 1 mm. divided into one hundred parts or 3 mm. divided into tenths and one-tenth divided in hundredths of a mm., etc. The stage micrometer is focused in the same way as a microscopic specimen and adjusted into position so that the rulings of the eyepiece micrometer are superimposed on them. It is then possible to evaluate the eyepiece rulings in terms of the standardized stage micrometer, after which the latter is removed and the specimen to be examined substituted in its place. Thereafter, the image of the eyepiece rulings will be superimposed on the image of the specimen and measurement can proceed. For precision work, a very accurately made eyepiece micrometer is used, a typical form of which has a thin glass plate upon which is ruled a cross and a double line. This is mounted on a slide immediately below a stationary micrometer scale and can be moved by means of micrometer screw. The cross is accurately set by the micrometer screw to coincide with the particle to be measured, the double line serves to count complete revolutions of the screw with the aid of the scale which is seen in the field of vision. The screw carries a drum which has 50 divisions and each division corresponds to a displacement of the cross through a distance of 0.01 mm. so that a complete revolution of the drum displaces the cross 0.50 mm. The actual readings of each interval of the drum head must be accurately calibrated for each objective by means of a stage micrometer. A group of instruments for use in micrometry is illustrated in Fig. 26.

Applications of Photomicrography

In closing, attention is directed to the photomicrographs comprising the Appendix of this paper, each of which was taken in connection with some definite engineering problem involving telephone apparatus. As the useful range of microscopic vision is extended farther and farther into the realm of higher magnifications, a more exact knowledge of materials is obtained and the effect is learned of physical and chemical forces acting to destroy or to build.

It has been conceded quite generally that about 1,500 diameters of magnification represents the limit of useful magnification. As previously stated this is a much disputed question. Laboratory studies, painstakingly carried out over a period of several years, have

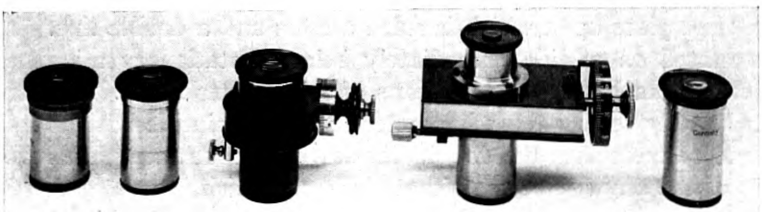


Fig. 26—Various types of eyepiece and stage micrometers used in connection with the microscope to obtain the dimensions of microscopic objects.

accomplished improvements in technique and in precision of adjustment of the equipment which have shown that remarkable resolution, depth of penetration and clearness can be attained in the case of metallurgical specimens, at extremely high powers. There seems little reason to doubt that our knowledge of metals can be augmented very materially by studies of their structures at high powers.

Moreover, it seems probable that the finest high-power objectives are of a quality beyond our ability to use them to best advantage because of our incomplete knowledge of how best to prepare specimens for examination at high powers.

It is impressive to evaluate magnification in terms more readily comprehended. For instance, the cross section of the average metallurgical specimen may be considered as a square whose side measures one-half inch. If we magnify this specimen 100 times, obviously we have an area measuring 50 inches on the side, but if we magnify it 10,000 times, then we have the equivalent of an area about 415 feet on a side or roughly, about four acres. An average picture at 6,000 diameters is 6 inches in diameter and therefore by a reverse order of reasoning, the actual area of the specimen under observation becomes 1/1000 inch in diameter.

APPENDIX

Fig. A. Meteoric iron consists of iron, nickel and the other elements usually found in steels such as carbon, sulphur, phosphorus, etc. The study of meteorites has contributed much valuable knowledge to the science of metallography. The Widmanstätten figures (shown by the arrangement of the constituents with reference to crystallographic planes) were generally considered characteristic of meteoric iron and it was believed that they were not to be found in manufactured iron and steel. Later this was shown to be an incorrect view.

(a) A meteorite which fell at Carthage, Tenn., containing 89.46% iron and 7.72% nickel and which shows the octahedral Widmanstätten structure. Magnification 4 X.

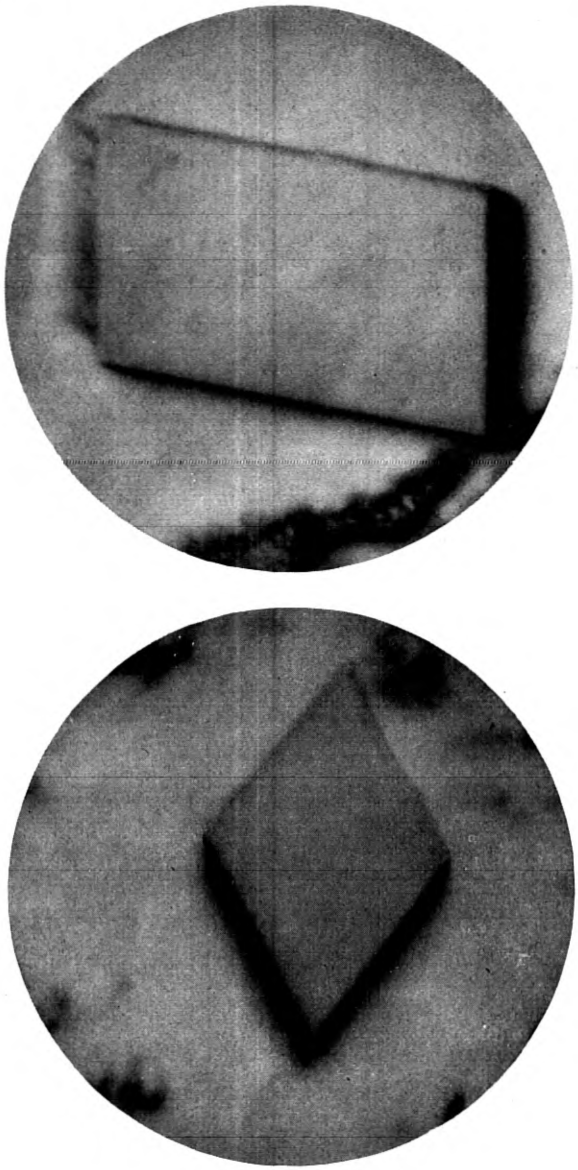


Fig. A

(b) Meteoric Crystals. The figures are sections through an octahedron and were developed by suitably etching a polished surface of the meteorite. Their perfect form is indicative of very favorable conditions of growth and is a corroboration of the octahedral crystalline form of the meteorite. Magnification 3500 X.

Fig. A

(c) A cast steel of 0.5% carbon in which the Widmanstätten or cleavage structure has developed somewhat similarly to that shown in the meteorite. The physical characteristics of the steel are dependent on the structural arrangement of its constituents, in this case pearlite (dark) and free ferrite (light). By suitable heat treatment this coarse structure may be refined and the physical properties of the steel greatly improved. Magnification 100 X.

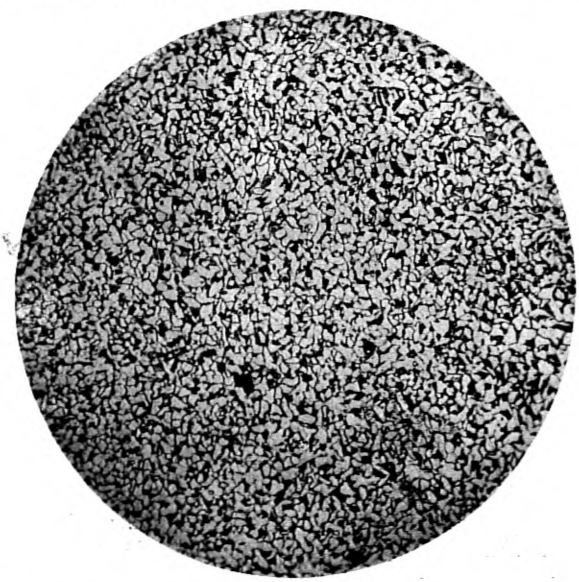


Fig. A

(d) The same steel as illustrated in "C" but after being refined (heated to 1000°C; air cooled; reheated to 650°C, and again air cooled.) Magnification 100 X.

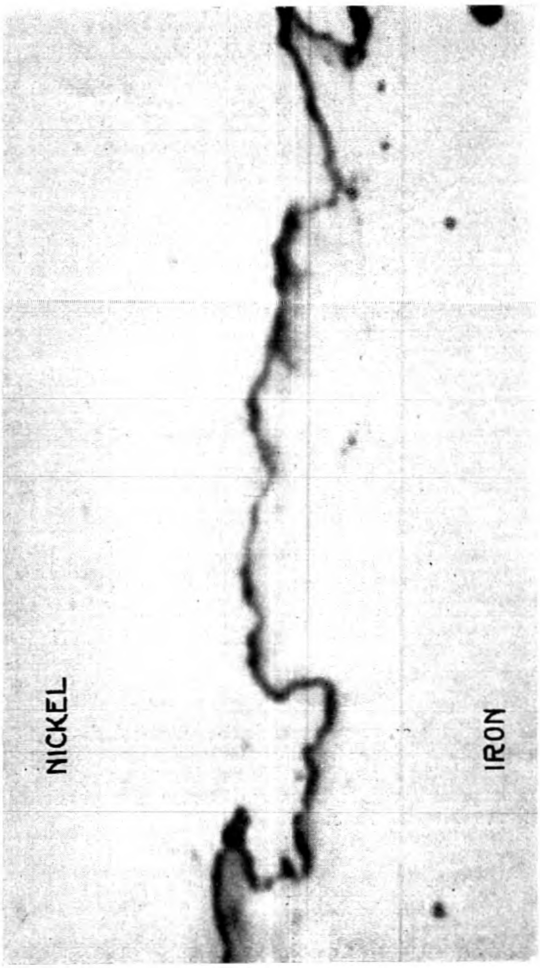
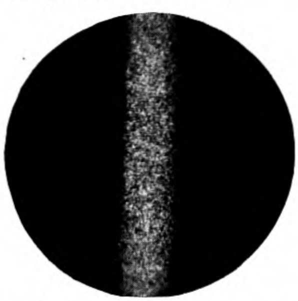


Fig. B. The application of high-power photography to the study of nickel finishes. One of the characteristics of an improved process for plating ductile nickel on iron is the interlocking or "keying" of the nickel and the iron Magnification, 6000 X.

Fig. C. Distribution of filler particles in soft rubber insulation as revealed by a transparent section. The section was cut in a microtome by "flashing" the rubber with liquid air which hardened it just sufficiently to cut properly. The specimen was photographed by polarized light and with selenite plates to secure contrast between the particles and the embedding rubber compound. Note agglomeration of the particles into large masses. The ideal condition of distribution would be attained when each individual particle is surrounded by rubber. Magnification 720 X.

Fig. D. Colloidal particles as seen through the ultra-microscope.
(a) Polymerized particles in a phenolic resin solution. Taken with the cardioid ultra-microscope and the Lucas Photomicrographic camera. Instantaneous exposure was necessary because the particles were in constant motion. Magnification, 220 X.



(b) Coloring matter in glass. The glass was colored saffranin and was transparent to the eye or with any other method of microscopic vision but with the slit ultra-microscope the colloidal coloring matter becomes visible. Also taken with the Lucas photomicrographic camera, a time exposure being necessary. Magnification, 100 X.

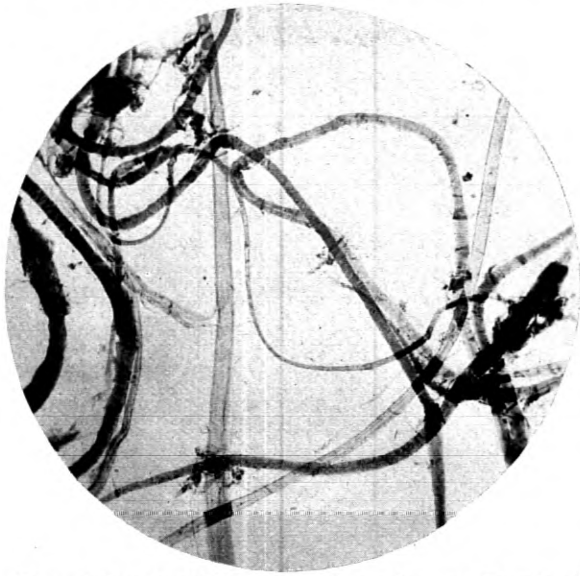


Fig. E. Paper fibers by 97 X. Note the surface markings: the gradation in color and the appearance of roundness possessed by some of the fibers. The photograph was taken with a modern medium-power apochromatic objective.

Microscopic examination of textile and paper fibers affords a means of identification second to none. The fibers are recognized by characteristics peculiar to each and by color reactions to different stains. Cotton, for example, appears as a flat ribbon-like fiber twisted spirally; linen is round and shows "joints" and cross markings. The specimen illustrated consisted mostly of linen with a small proportion of cotton added.

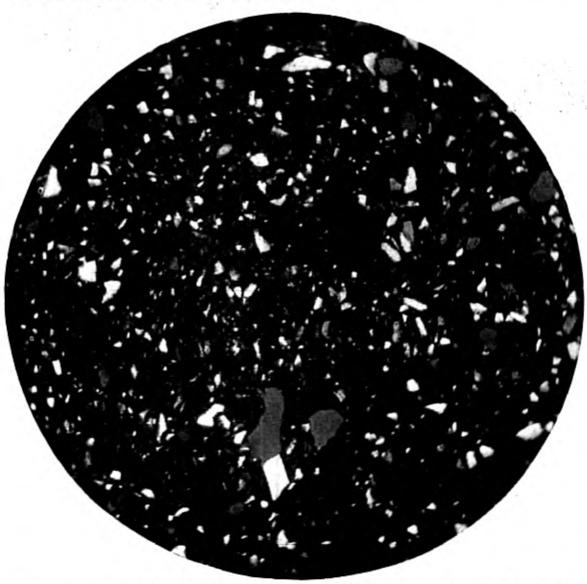


Fig. F. Electrical porcelain by polarized light, magnification 100 X.

The quality of the porcelain may be judged to a considerable degree by a microscopic examination. The degree of vitrification is indicated by the rounding of the sharp corners on the quartz grains; whether or not the porcelain is homogeneous may be determined by the uniformity in distribution of the undissolved particles, and fissures, cracks, or voids are readily seen. All of these factors influence the physical characteristics of the porcelain.

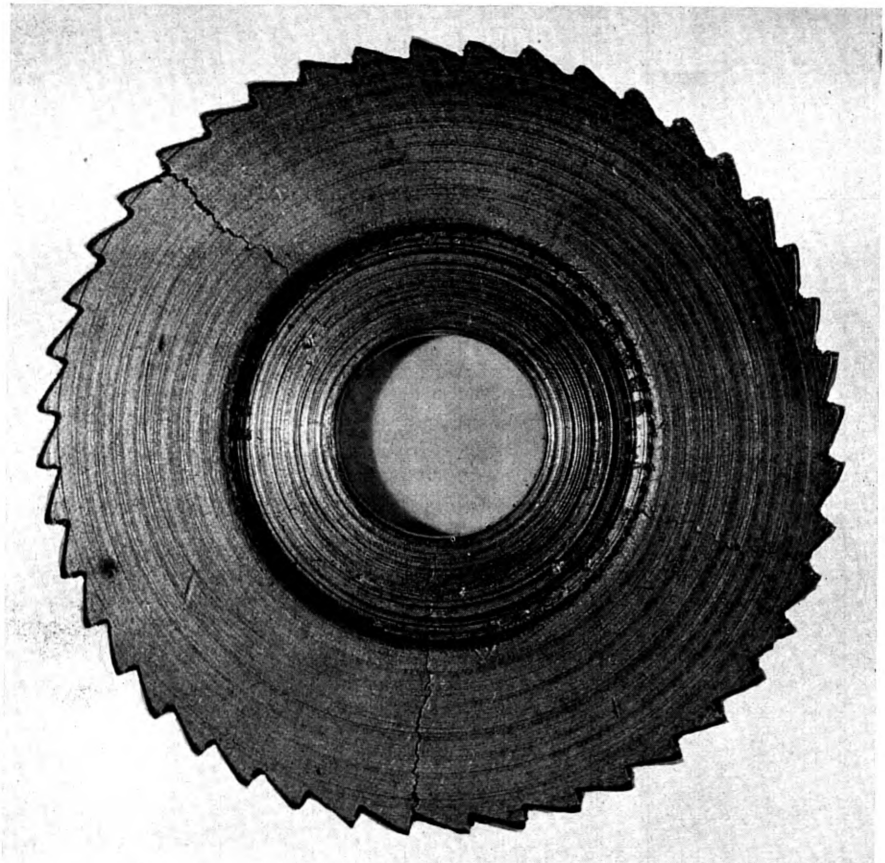


Fig. G. Season cracking of aluminum bronze ratchet wheels.
(a) Ratchet wheel at magnification 2¾ X.

Fig. G

(b) Showing a large crack, at magnification of 23 X.

Fig. G

(c) Intercrystalline nature of the cracks and the severely worked condition of the metal as indicated by several groups of slip bands traversing each crystal grain.

These ratchet wheels developed radial cracks while in storage or in service. Some of the cracks were so large as to be plainly visible to the unaided eye and others were of microscopic dimensions. They resulted from the metal being severely cold-worked at the time the parts were machined and then left in a strained condition. The intercrystalline nature of the cracking is shown in "c" which is characteristic of season cracking. This illustration also shows the crystal grains traversed by several groups of slip bands, indicating the severity of the cold-working.

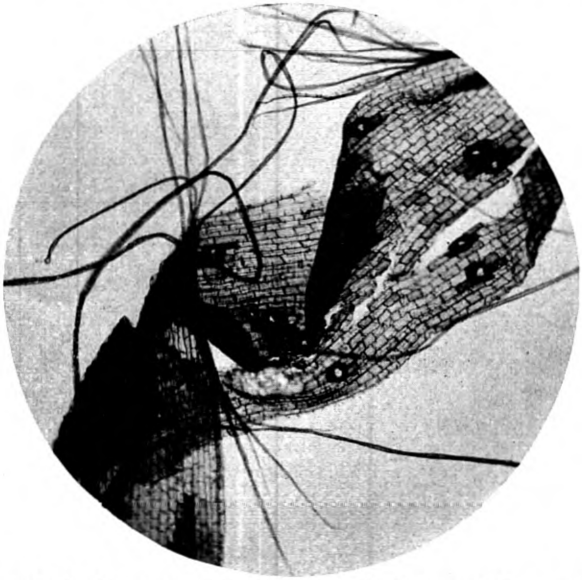
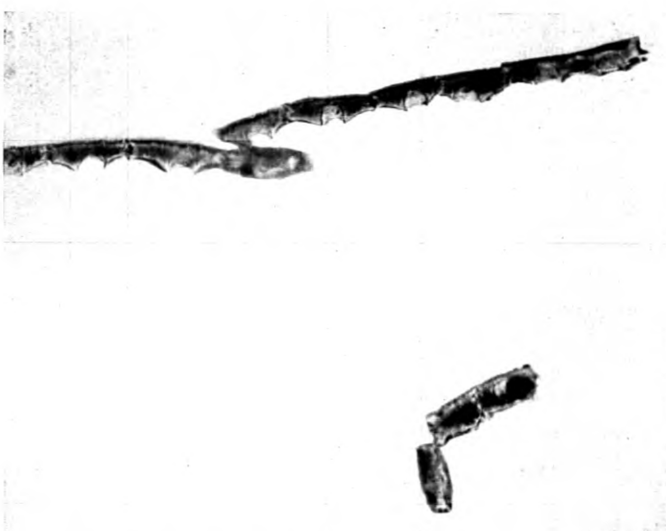


Fig. H. Manila hemp rope is used extensively in telephone work and the fiber from old rope is used in paper for cable insulation which finds its way into the plant.

Microscopically the fiber is identified by certain characteristics, prominent among which are the silicified tabular cells known as stegmata. If the fiber is burned and treated with dilute acid the stegmata remain behind, resembling strings of beads.

Manila hemp makes the best cordage but it is somewhat difficult to distinguish the fiber from that of sisal which produces inferior cordage. The presence of the silicious skeletons of the stegmata and the color of the ash (grayish-black in the case of Manila hemp and white in the case of sisal) aid in the identification of the fiber.

(a) Manila Hemp Fibers, magnification 50 X.



(b) Ash of Manila Hemp, magnification 450 X.

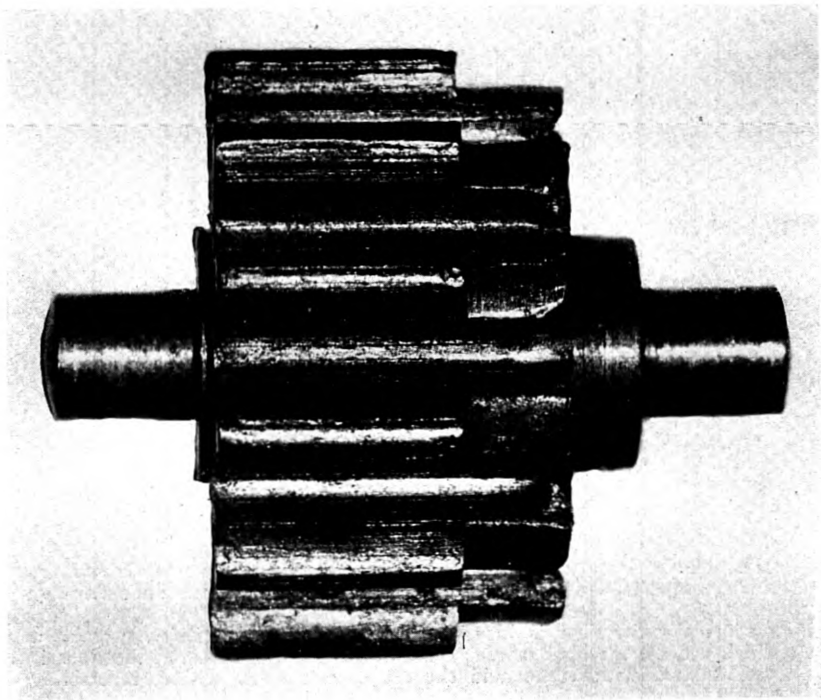
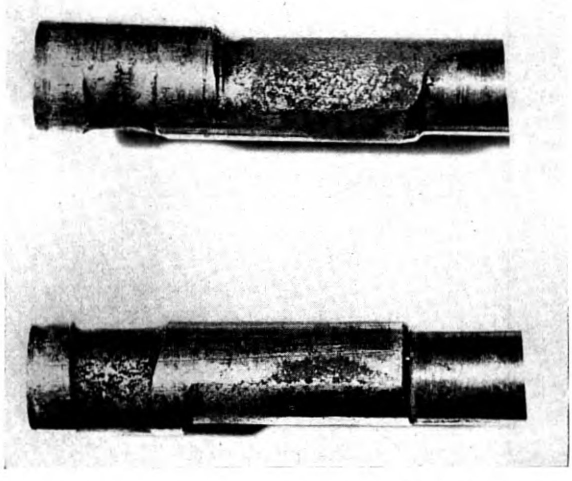


Fig. I. A further illustration of low-power photomicrography in the study of telephone parts:
(a) Intermediate pinion of calling dial. Diameter of pivot .050 inch. Magnification, 14 X.



(b) The effect of laboratory wear tests on small shafts. Magnification, 4 X.

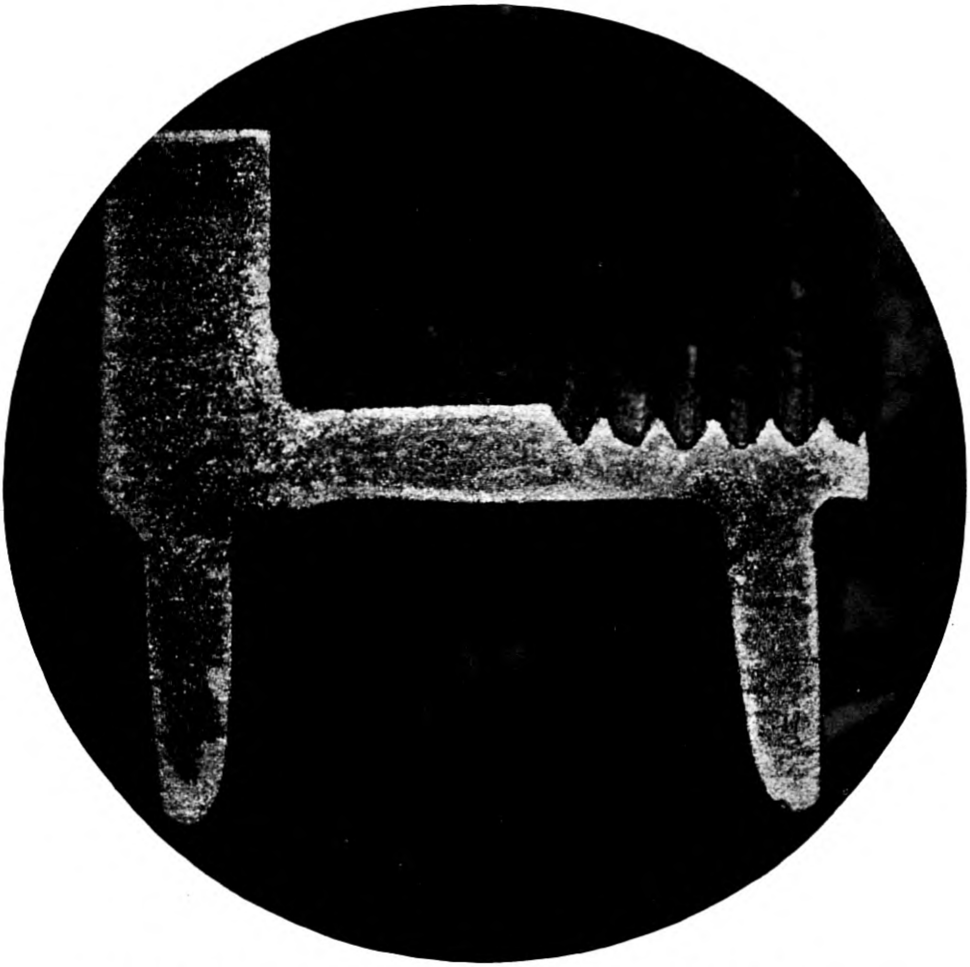


Fig. J. (a) Faulty pack hardening of car wheels for toll ticket distributing system. The object of pack hardening is to impart a highly carburized wearing surface to the otherwise soft steel part. The interior remains soft and ductile. Lack of uniformity in hardening or insufficient depth of the carburized zone causes soft spots which result in unequal wear. The magnification is 11.2 X.

Fig. J

(b) Showing inappreciable depth of carburized zone, and a large non-metallic inclusion in the steel. Inclusions of this sort denote poor quality or dirty steel. Magnification, 100 X.

Fig. K. Steel of 1.5% carbon heated to 825° C. and quenched in oil. This medium-power photomicrograph at 100 X really tells very little about the steel, except that it possesses a fine structure.

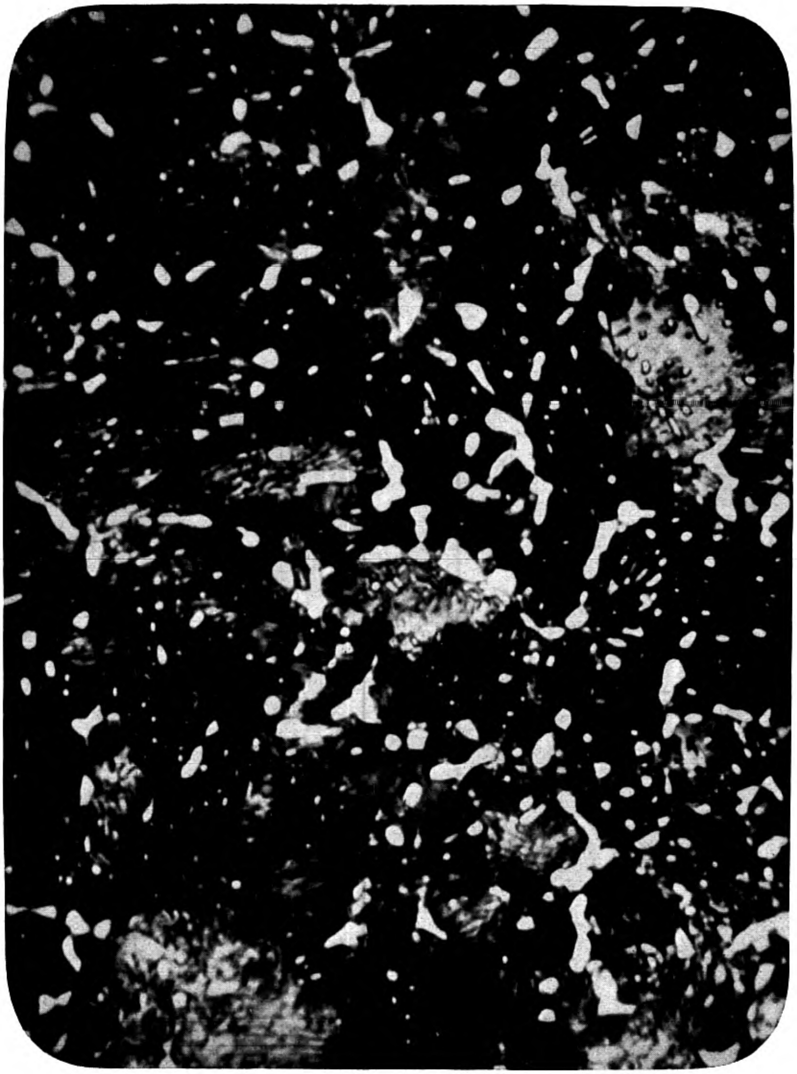


Fig. K. (*continued.*) At 2170 X the specimen is seen in the process of being converted to spheroidized cementite. The cementite (iron-carbide) which is the light constituent is in the process of transforming from a laminated form to small globules. Grain boundaries are still marked by accumulations of cementite but this is spheroidizing. In the light patches stratification of Ferrite and cementite is just visible.

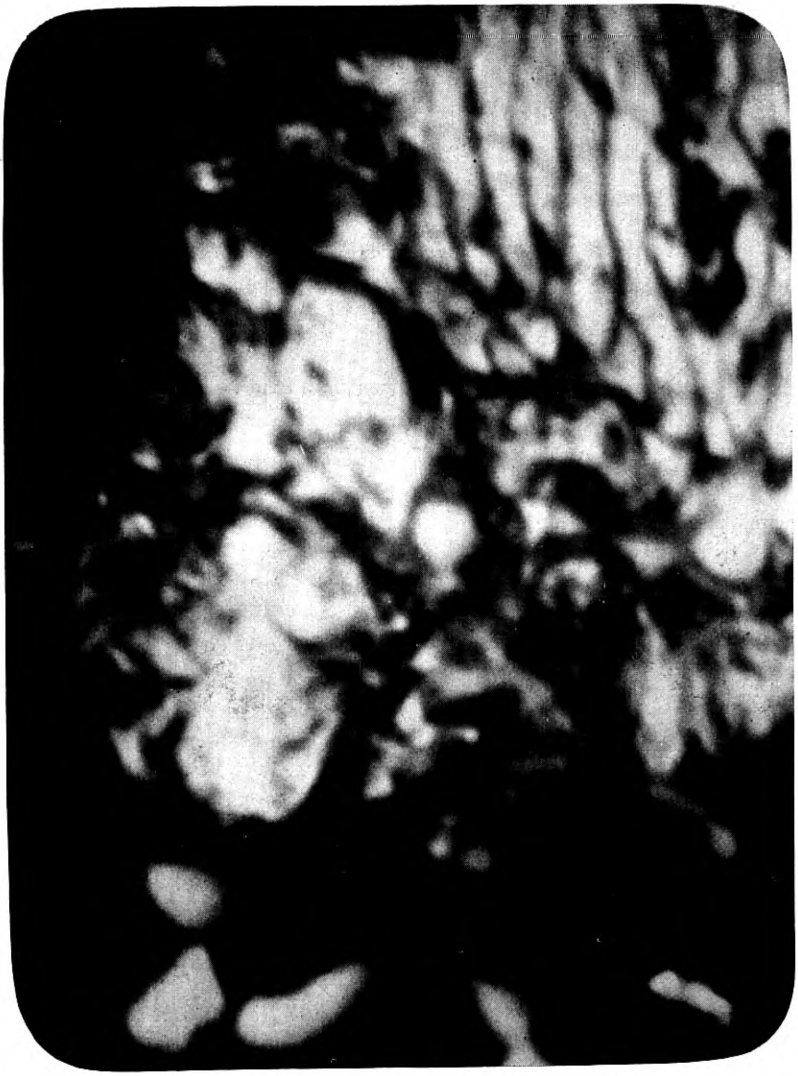


Fig. K. (*continued.*) Under higher magnification one of these patches shows clearly the remaining vestige of laminated structure and the commencement of spheroidization. Magnification 9000 X.

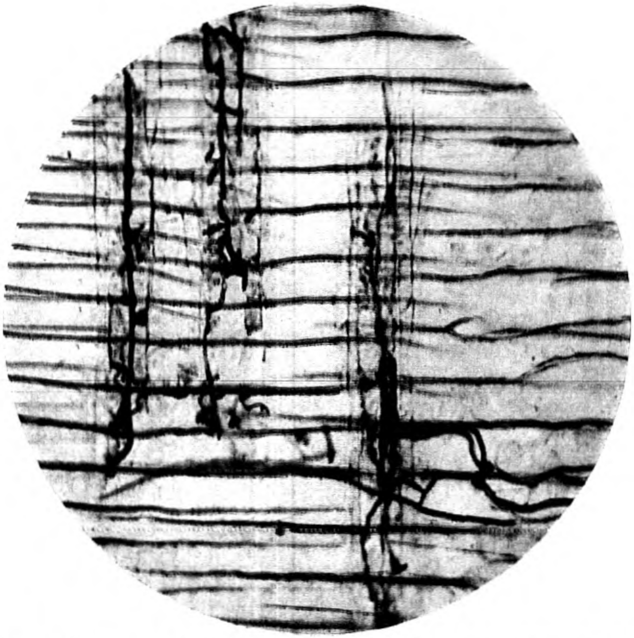


Fig. L. Direct autochrome color reproduction of a stained specimen of Southern Yellow Pine showing sap-stain fungus mycelium. Magnification 100 X.
This particular fungus is harmful to the extent that it causes a discoloration of the sapwood, which assumes a blue color in place of the usual straw-yellow. Wood-destroying fungi differ somewhat in their appearance from the one illustrated.

Fig. M. Direct color photomicrography by the autochrome process of a radial section of mahogany wood. Magnification 50 X. Mahogany is one of the best of cabinet woods and finds wide application in the telephone plant.

A Clock-Controlled Tuning Fork as a Source of Constant Frequency

By J. G. FERGUSON

NOTE: The art of electrical communication employs such a wide variety of methods for the transmission of intelligence that it utilizes alternating currents whose frequencies cover the entire range between a few cycles per second and several million. With the increasing use of these methods, it becomes more and more imperative that determinations of the frequency of any alternating current may be made with extreme accuracy. In particular, recent developments in carrier current telephony and telegraphy over wires have placed exceedingly rigorous limits on the frequency adjustment of certain types of apparatus. It is many times necessary to hold such equipment as oscillators or filters to within 0.1 per cent. of given frequency values under commercial operating conditions. This means that calibrating devices used in the manufacture and maintenance of such circuits must be reliable to 0.01 per cent. and that the primary standard should be good to about 0.001 per cent. or one part in 100,000.

The present paper discusses one of the methods recently developed in the Bell System Laboratory for obtaining a source of practically constant frequency with which other frequencies may be compared. It consists of a clock-controlled tuning fork making 50 vibrations per second and, as is shown, the maximum deviation of its frequency from the mean is less than one part in 50,000.

A study has also been made of means for improving the constancy of the control clock and a new type of clock mechanism consisting of an electrically actuated pendulum, the impulse of which is controlled by a photo-electric cell, is suggested.—EDITOR.

INTRODUCTION

THE art of clock making is of such long standing that there have been few improvements of note in the last fifty years tending to increase accuracy. The average rate of oscillation of a good clock when taken over a sufficiently long period of time as, for instance, a day, can be held constant to about one part in 1,000,000. This accuracy is sufficiently high for all ordinary requirements in the measurement of time, including the field of electrical communication.

However, in electric measurements, the problems which present themselves ordinarily require the accurate measurement of intervals very much shorter than a second which is usually the smallest interval registered by the average clock. In solving these problems, we are therefore forced to the alternative either of designing a clock to have a period very much shorter than those of existing clocks or of using some form of short period oscillator whose uniformity can be controlled by the second impulses from a clock.

The first method has been admirably worked out as described by other members of the staff of this laboratory.[1] In this system a

[1] Paper by J. W. Horton, N. H. Ricker and W. A. Marrison, presented at the annual convention of the American Institute of Electric Engineers, June, 1923.

hundred cycle fork is kept in constant oscillation by a regenerative method, the conditions being so controlled that the mean period of the fork compares favorably with that of a good clock.

The attraction of the second method lies in the possibility of obtaining a sufficiently constant standard of frequency with nothing more than a good clock and standard auxiliary apparatus easily capable of application to any oscillating system. Such an outfit could be made available in cases in which the expense incident to the installation and maintenance of more elaborate equipment would not be justified.

Requirements of a Clock-Controlled Frequency Standard

It is a comparatively simple matter to control or operate a fork, or other oscillating system, by means of periodic impulses from a clock, so that the total number of oscillations will be some definite multiple of the number of impulses from the clock. However, the present requirements are more severe than this. It is necessary to have the oscillator operated so that each oscillation will be sensibly equal in magnitude and duration to every other oscillation. In other words, it is not sufficient that the clock and the oscillator keep in step over a given period of time, but the instantaneous frequency of the fork must not depart appreciably from the mean frequency. This requires a form of control which will not be to any extent discontinuous, but which will change uniformly in proportion to the divergence of the oscillator from the clock. Such a form of control in turn requires that the frequency of the oscillator itself be sufficiently constant when uncontrolled, to reduce all momentary fluctuations and rapid frequency changes to a minimum. This requirement is best satisfied by an oscillating system having a low decrement. Since a mechanical system is usually far superior to an electrical system in this respect, and since the most available mechanical oscillator for the range of frequency in question is a fork, our choice naturally falls on this form of oscillator.

A good fork maintained in continuous operation by some electrical means, such as regeneration, or a make and break contact and a driving magnet, is a comparatively simple system and is capable of a high degree of constancy.[2] It therefore satisfies all of the requirements for our purpose, but there remains the devising of some control which will be proportional to the divergence of the fork from the clock controlling it. In order to use any such control it is practically necessary to integrate the oscillations of the fork so that we may obtain a

[2] H. M. Dadourian, Phys. Rev. 13, page 337, 1919, "On the Characteristics of Electrically Operated Tuning Forks."

time interval equal to the number of cycles of the fork which we desire to make equal to the time interval between successive clock impulses. This is readily accomplished by means of a phonic wheel or synchronous motor operated by the fork. This motor may be connected to any form of gear train in order to get the necessary integration.

The requirements so far outlined do not limit the frequency of the fork in any degree except that we must be able to integrate its periods, and if a mechanical means is used as outlined, this probably sets an upper limit on the frequency at 400 or 500 cycles. However, practical considerations will generally make the most satisfactory frequency considerably lower than this, since it is an easier matter to compare unknown frequencies with a low frequency standard rather than with one of high frequency.

Method of the Control of the Fork by the Clock

The fork used in the system described below is of the same type as that tested by Dadourian. It is operated by a driving magnet and make and break contact, and was originally designed for use in multiplex printer telegraph circuits. It can be adjusted to operate at 50 cycles and is designed to drive a synchronous distributor which rotates once for every 10 cycles of the fork. By means of a 5 to 2 reduction gear and a contact operated by it, an impulse may be obtained once every 25 cycles of the fork. If the fork oscillates at exactly 50 cycles per second, the time interval between the impulses will be exactly one-half second, and this time interval will be shorter or longer, according as the speed of the fork increases or decreases.

The control system used is designed to affect the frequency of the fork in proportion to the difference between half second intervals as measured by the clock and the time required by the fork to complete 25 cycles. Fig. 1 shows the details of this control. Fig. 2 is the schematic diagram. Referring to Fig. 2 the contact marked "Fork" is the contact obtained every 25 cycles from the fork and the contact marked "Clock" is that obtained every half second from the clock. Each of these contacts is adjusted to remain closed for a period of approximately .05 second when operated.

The control operates as follows. When the clock contact closes, the relay operates and locks until the fork contact closes and short-circuits the winding of the relay which then releases. During the time that the relay is operated, the condenser C is charged through the resistance r_1 by the battery B_1. The voltage of this battery is such that when applied to the grid of the vacuum tube, it will just reduce the space current to zero. The condenser C continuously discharges through the resistance r_2. The mean potential on the condenser is thus applied to the grid of the vacuum tube and modifies the space current, which, in turn, is passed through the damping coil

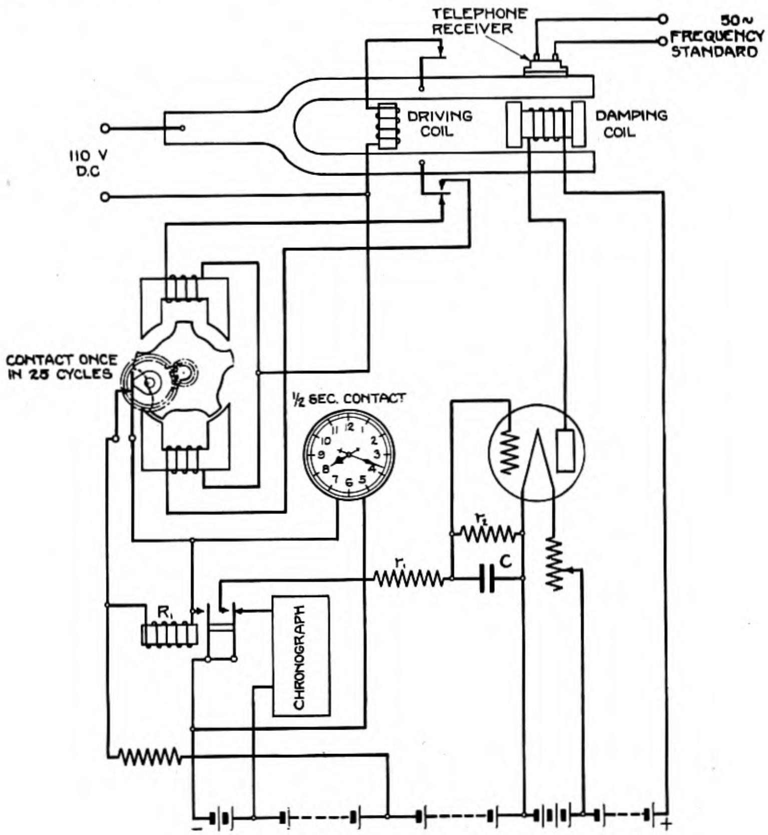


Fig. 1—Clock Control of Fork Frequency

of the fork. A stable condition is reached when the condenser discharge each second is exactly equal to the charge. Any variation in the condenser potential varies the current through the damping coil and changes the frequency of the fork. Now if the period, during which the relay remains operated, increases, the mean potential on the condenser will gradually increase. This will increase the mean negative grid potential, reducing the mean space current through the tube and through the damping winding, thus reducing the damping on the fork and increasing its frequency.

This method of control is slow yet sensitive to very slight changes in frequency.

The method of controlling the frequency of the fork by a damping winding was found to be the most simple and satisfactory method. The amount of variation of frequency which this winding will produce under extreme conditions should be slightly greater than the maximum variation to which the fork is subjected in operation when uncontrolled. This has been found to be about .05%, when temperature

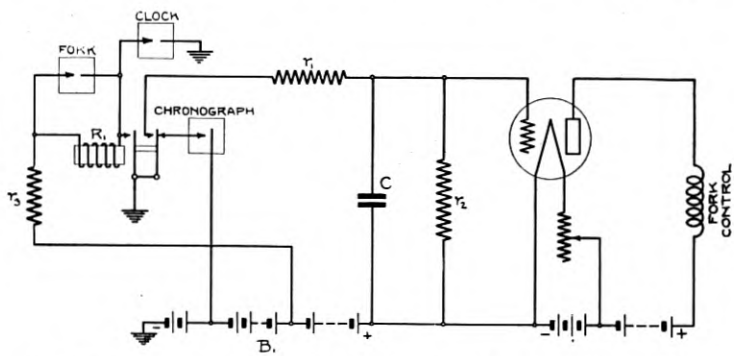


Fig. 2—Clock Control of Fork Frequency. (Schematic Diagram)

variations are held to within a few degrees. The only requirement for the coil is that the current available from the vacuum tube must be sufficient to produce the necessary control. This requirement is easily met. It has been found possible with the equipment described to obtain an effect at least 10 times greater than necessary.

The change in frequency of the fork due to the current in the damping winding is a combination of several effects. The current will increase the decrement of the fork, due to the losses induced in the metal of the fork while vibrating in a magnetic field. This will cause a decrease in frequency. The magnetic force acting on the tines of the fork, even though it be assumed to cause no losses in the fork, is unsymmetrical, having a greater effect at the ends of the swing of the fork. This unsymmetrical force may also be shown to cause a decrease in frequency. Again there is a change in frequency due to the change in amplitude alone. For the type of fork here used this change may be an increase or decrease, depending on the range over which the change occurs.

It is obvious that when the control is operating, the voltage of the condenser and hence the space current of the vacuum tube, will fluctuate each half second. Since it is only the mean value of space current that is used to control the fork, it is important that this fluctuation be reduced to a small amount. This may be done by using a large capacity C or a large resistance R_1. However, the effect of increasing the capacity or resistance is to increase the time required for the control to change, when compensating for changes in the fork frequency. Accordingly the values chosen must be a compromise. If we assume that the control is capable of giving a maximum change in frequency of .1%, and we allow a fluctuation in this control of 5% each half second, this will cause a fluctuation in frequency of 5%, of .1%, or .005%. However, the inertia of the fork prevents it from following such a rapid fluctuation in damping current and hence the actual change in fork frequency is very much smaller than just indicated.

The fact that non-cumulative fluctuations in the control as great as 5% have only a negligible effect on the fork frequency is an important point. Such fluctuations are likely to arise through hunting in the synchronous motor, irregularities in the time of operation of the relay, etc., and since their effects average one another out, there is no danger of their being transferred to the fork.

The ratio of the charging resistance to the discharge or grid leak resistance is not a governing factor, except that the charging resistance must be less than the discharge resistance. The phase position of the fork to the clock under normal conditions is also governed by the relative values of these resistances. For the present circuit r_1 has a resistance one-half that of r_2, and these resistances and the condenser are of such values that it takes approximately 15 minutes for the fork to come into the correct phase relation with the clock when started under the most unfavorable conditions.

While this method of control will hold the fork frequency for an indefinite period in synchronism with the clock, it is possible that the phase relation of the clock to the fork may change. This change may be periodic, that is, it may take the form of an oscillation about the mean phase position, or there may be a gradual change due to changes in the various constants of the control occurring over comparatively long intervals. For instance, any change in the ratio $\frac{r_2}{r_1}$, such as might occur with temperature, will change the phase relation between the fork and the clock.

Chronograph records show that there are no phase changes greater than one cycle of the fork over periods as large as 8 hours. To determine the possibility of hunting, that is, of oscillation of the fork frequency around its mean value, the phase relation was actually disturbed and a chronograph record taken of the readjustment. This will give the period of the oscillation, if any, and the amount of damping.

Fig. 3 shows one of these records. The chronograph was connected in the circuit as shown in Fig. 1 and a record was taken over a period of about 20 minutes after starting the fork. This record shows the length of time in each half second that the control relay was operated. At starting this period is about .11 second. After about 8 minutes it becomes a maximum equal to .2 second and there is no appreciable change over the next 5 minutes, showing a permanent condition has been reached. Accordingly we may conclude from this record that any oscillation about the mean value of the control

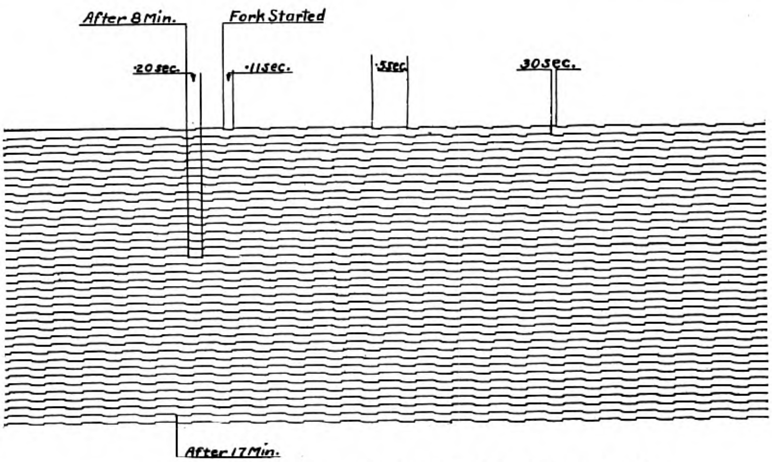


Fig. 3—Chronograph Record of Fork After First Starting

is almost if not quite critically damped. Several other records taken with even greater phase displacements bear out this conclusion. This practically precludes hunting after the phase angle has been once adjusted.

Accuracy of the Clock-Controlled Fork

The accuracy of the fork has been checked in two ways. For long periods of time, chronograph records have been taken at intervals over a period of 8 hours and the maximum variation of the fork from the clock in this period has been found to be less than .02 second, or one cycle. Smaller periods of time cannot be measured accurately on the chronograph used. If we are dealing with periods of time of more than 15 minutes, this gives an accuracy as high as one part in 50,000.

For small time intervals, an entirely different method for measuring the constancy of the fork must be used. Two methods are available. We may either compare the high harmonics of the fork directly with some high frequency which can be held extremely constant over short periods of time, and observe fluctuations in the relative values of these frequencies, or we may compare the fundamental frequency of the fork with a high frequency by some means which will enable us to measure the divergence from an exact integral multiple relation in terms of the higher frequency.

To explain in more detail, we may pick out the one hundredth harmonic of the fork by means of a filter and amplifier and compare it with a 5000-cycle frequency obtained from a constant frequency oscillator by some method of detection which will allow us to count the difference in cycles. By this means we may observe variations in the relative rate of the fork and the oscillator to an accuracy of about one-tenth of a cycle over a period of a few seconds, and this gives us a comparison to an accuracy of 1 part in 50,000. The principal objection to this method is the difficulty involved in separating the higher harmonics of any alternating current wave obtained from the fork. For instance, the separation of the hundredth harmonic from those immediately above and below it would require a circuit so selective that it would probably be very difficult to construct and cumbersome to operate.

If we had means to determine when some high frequency such as 5000 cycles was an exact multiple of the 50 cycles and to measure the difference in terms of the 5000-cycle wave, we would be able to obtain the same results, and avoid the above difficulty.

A device which will allow us to do this is the low voltage cathode ray tube developed by Johnson[3]. The two frequencies to be compared are connected to the two pairs of plates of the tube and the combination of the two deflections causes the luminous spot to trace out a path which repeats itself indefinitely if one frequency is an exact integral multiple of the other, and a stationary figure is produced. In this way any frequency which is a multiple of the fundamental 50 cycles may be accurately determined. As the method of comparison is an electrostatic one practically no power is used.

For the type of tube used, a deflection of about 1 centimeter is obtained for a potential difference of 10 volts between plates, and

[3] J. B. Johnson, *Bell System Technical Journal* Nov. 1922, "A Low Voltage Cathode Ray Oscillograph."

frequencies having ratios as high as 100 to 1 may be readily compared. For ratios of the order of 100 to 1 the lower frequency is preferably stepped up to a high voltage to give an equivalent deflection of as much as 25 centimeters, thus giving a spacing between cycles for the high frequency of approximately 0.5 centimeter. Of course, the whole 25 centimeter deflection is not shown on the screen but this is unnecessary. The value of the ratio cannot be at once determined by this means, there being no appreciable difference between the figure for a ratio of 100 to 1 and 99 to 1, but this ratio may be readily determined by comparing each frequency separately with an intermediate frequency such as 500 cycles.

Having determined the ratio between the high and low frequencies, it is possible, by drawing a reference line across the screen, to determine whether or not they are keeping step with one another. Thus for a comparison of 50 cycles against 5000 cycles, if we get a motion of 2 waves in 10 seconds, this represents a deviation from exact synchronism of 2 parts in 50,000.

Comparisons made in this way between the 50-cycle fork and a vacuum tube oscillator giving a constant frequency of 5000 cycles show no deviation in the mean period of the fork greater than 1 part in 50,000 for observations extending over several minutes. If deviations greater than this were observed, they might equally be ascribed to the auxiliary oscillator but the fact that they do not occur means either that the fork is constant to better than 1 part in 50,000 or that both frequencies vary in exactly the same way which is very improbable.

The above method of comparison does not require a sine wave of current from the fork. In fact it has been found advantageous to have a somewhat distorted wave since an unsymmetrical figure on the luminous screen of the tube is more easily observed. This is due to the fact that one-half of the figure moves across the screen in one direction while the other half moves in the opposite direction. In order not to confuse one half with the other, it is highly desirable that they be dissimilar in shape and this is accomplished by using a distorted wave as the lower frequency. Sufficient distortion is secured by mounting an ordinary telephone receiver in close proximity to one prong of the fork as shown in Fig. 1 and amplifying the e.m.f. thus obtained as much as necessary to obtain the desired voltage.

By means of the simple control system described above, it has been possible to obtain a fundamental frequency so free from fluctuations as to be constant over short or long periods of time to approximately one part in 100,000.

Accuracy of the Clock

So far we have not considered the possibility of error in the clock as a factor. Of course, the fork cannot keep better time than the clock which controls it.

The clock used at present was made by L. Leroy and Co., Paris, electrically driven and beating half seconds. The drive consists of an electric circuit including a single primary cell mounted in the clock, a driving coil and a contact which is closed by the escapement wheel for approximately .1 second in each second. Attached to the lower end of the pendulum is a steel bar which moves into the driving coil as the pendulum oscillates. The electrical impulse is so timed that the driving coil gives the pendulum a slight pull as it is entering the coil. This impulse is sufficient to keep the pendulum oscillating. An additional contact on the clock is used to furnish an electrical impulse for timing purposes.

Time records of the clock have been kept over a period of several months and the rate has been found to be constant to about one-half second a day, which is better than 1 part in 150,000. Since this accuracy is not very much greater than the precision with which the fork keeps in step, any further accuracy will require refinements in the clock itself. With this object in view, an investigation was made of the possibility of obtaining greater accuracy from the existing clock.

Errors are of two kinds. First, if the timing contact is obtained by the operation of the escapement wheel, there may be a cyclic variation in the length of time between successive impulses extending over one revolution of the wheel, (1 minute) even though the pendulum keeps perfect time. This has been found to be the case in some of the best clocks in the country. This error can be overcome by taking the contact direct from the pendulum. The contact we are using at the present time is of this type obtained from the pendulum by means of a photo-electric cell.

The optical system is shown on Fig. 4. Light from the source A is concentrated on the mirror, which in turn reflects it on to the photo-electric cell. When the pendulum passes through the center of its stroke, it momentarily cuts off this beam of light. This causes a large increase in the resistance of the photo-electric cell, the change taking place almost instantaneously.

Referring to the diagram of connections on Fig. 4, the potential of battery B is divided almost equally between the photo-electric cell and the grid of the tube if the grid leak is made approximately equal to the resistance of the cell when exposed to the light. This gives a negative potential to the grid sufficient to cut off all space current, and the relay R_2 remains unoperated. When the pendulum cuts off the light to the photo-electric cell, the resistance of the cell rises immediately and the grid voltage drops to a very small value. Enough space current will pass now to operate the relay R_2 and a

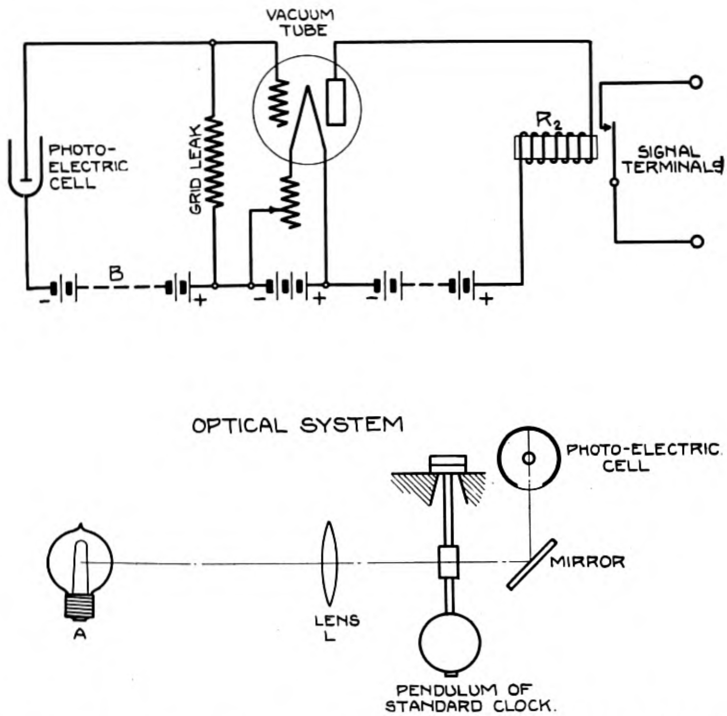


Fig. 4—Circuit for Obtaining Electrical Impulses from Standard Clock Using Photo-electric Cell

signal is transmitted of the same duration as the time the light is cut off the cell by the pendulum. There is no appreciable time lag in the photo-electric cell or vacuum tube.

The principal requirement in setting up this circuit is to obtain a vacuum tube having a resistance between filament and grid including wiring, which is under all conditions considerably greater than the minimum resistance of the photo-electric cell. If this resistance drops much lower, the circuit becomes inoperative even though no additional grid leak is used.

The only irregularity introduced in this system is in the operation of the relay, and as this is a fast operating relay this error will be less than the accidental irregularities in a contact obtained from the escapement wheel even excluding errors due to eccentricity.

This method of obtaining an electrical impulse from a clock is of great value as it may be applied to practically any clock which may not have any other method of producing impulses.

The second type of error is due to variations in the rate of the clock. Two fundamental requirements in the design of an accurate clock are that the impulse delivered to the pendulum be symmetrical about the mid-point of its swing and be not subject to irregularities in magnitude or duration, and that the pendulum be free at all other parts of its swing. These requirements are fairly well met in the

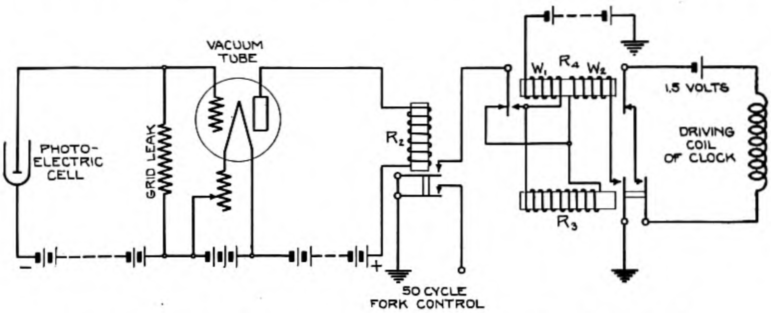


Fig. 5—Circuit of Photo-electric Cell Drive of Standard Clock

present clock. However, the magnitude of the impulse depends on the constancy of the voltage of the driving cell which is a single primary cell of rather small size, and the duration of the impulse may be somewhat variable due to the possible eccentricity of the escapement wheel and due to the method of operation. The pendulum too, is not entirely free from constraint at any part of its swing. These errors may all be avoided or at least considerably reduced by the use of the impulse obtained from a photo-electric cell to drive the clock and by the use of a more constant source of primary voltage.

The use of this type of drive has accordingly been investigated in connection with this clock. It is obvious, since the driving impulse is one of attraction between the coil and the bar carried by the pendulum, that it must be exerted only once per second, that is, when the pendulum is entering the driving coil and not when it is returning. The circuit used is shown on Fig. 5 and operates as follows:

When the relay R_2 operates the first time in the second, it closes the circuit through the winding W_1 of the relay R_4 and through relay R_3. This operates the relay R_3 and closes the circuit through the driving coil of the clock. The current through the one winding of

R_4 is not sufficient to operate it. As soon as the relay R_2 releases, current will pass through all the windings on both relays which in turn closes the relay R_4. This opens the circuit through the driving coil of the clock. The impulse given to the pendulum is, therefore, the duration of the operation of the relay R_2, or the time during which the light is cut off the photo-electric cell during the swing of the pendulum to the left. When the pendulum swings to the right and the relay R_2 operates, R_3 is short-circuited and releases, R_4 being held up by winding W_1. When R_2 releases, it releases R_4 bringing the circuit back to normal. Since the circuit through the driving coil of the clock is closed only when the relay R_3 is closed, and the relay R_4 is released, there is only one impulse per second given to the pendulum.

During a period of operation by this method covering several days the clock gave as satisfactory performance as with the mechanical drive, but while the present gear train is connected to it, no appreciably better performance can be obtained than at present, and accordingly it is proposed to carry out further work along this line with an experimental pendulum having no mechanical connections. By using a good compensated pendulum and mounting it suitably in a constant temperature hermetically sealed case, it appears probable that a photo-electric cell drive would produce a more constant rate of oscillation than the best clocks of existing types. The advantage of this type of drive over other types is the fact that the pendulum is absolutely free from all mechanical constraint at all parts of its swing. The problem of supplying an uninterrupted current for the light and power could readily be solved by the use of duplicate apparatus.

The general method outlined in this paper for synchronizing a fork with a clock has a very wide field of usefulness, and is not limited to the particular application described. For instance, in place of the clock we may substitute another fork and distributor, and we are thus enabled to hold 2 forks with their distributers in exact synchronism by means of an impulse transmitted at a constant time interval of about once every half-second.

By substituting the field coils of a motor for the damping winding on the fork, we are able to hold the speed of the motor in synchronism with the clock, the only requirement being a step down gear on the motor to furnish the desired contact.

The general principle involved is not dependent on the use of a vacuum tube, and if other means of control based on this principle be adopted, very large powers may be controlled in the same way.

Some Contemporary Advances in Physics—II

By KARL K. DARROW

NOTE: Dr. Darrow, the author of the following article, has made it a practice to prepare abstracts and reviews of such recent researches in physics as appear to him to be of special interest. The results of Dr. Darrow's work have been available to the staffs of the Bell System laboratories for some time and having been very well regarded, it is thought that such a review, published from time to time in the TECHNICAL JOURNAL, might be welcomed by its readers.

The review cannot, of course, cover all the published results of physical research. The author chooses those articles which appear significant to him or instructive to his readers, without attempting to pass judgment on the scientific importance of the different papers published. It is not intended that the review shall always assume the same form; at one time it may cover many articles, at another be devoted to only a few, and it may occasionally treat of but a single piece of work.

The present installment, which is Number II, is devoted very largely to the subject of atomic structure.—EDITOR.

WE know quite definitely that an atom consists of a massive positively-charged nucleus with a certain number of electrons in its vicinity; but of the arrangement of these electrons in the strict geometrical sense we know very little—indeed, we do not certainly know even whether they are in motion or not. Apparently there are many possible arrangements for each kind of atom; one of these is a permanent arrangement, in the sense that when once established it is not changed so long as the atom is not disturbed from outside; the others are transient. In addition to the arrangements of the electrons in the neutral atom, there are the arrangements of the remaining electrons when one or more of the normal quota are lacking. When an atom changes over from one of these arrangements to another, it must take in or give out a definite quantity of energy. Another way of saying this same thing is that to each distinct arrangement of the electrons there corresponds a distinct value of the energy of the atom. These values of the energy of the atom are directly or indirectly measured, often with great precision; they are the data of experiment. The very precise statements, or at all events very definite statements, which are frequently made about the "structure" of the atom, usually refer only to these energy-values and the relations between them.

The simplest question that can be asked about the arrangement of the electrons is, whether they all occupy identical positions—being, for example, evenly distributed over the surface of a sphere or the circumference of a circle, with the nucleus at its centre. If this is true, the same amount of energy will be required to remove any electron from the atom as to remove any other. In the extreme opposite case there would be as many different amounts of energy required to remove an electron from the atom as there were electrons. Now, when radiation of a definite frequency ν falls upon a group of atoms, any particular atom will either ignore the radiation, or else will absorb a definite quantity of energy $h\nu$ from it. (The letter h, as usual, denotes Planck's constant, $6.56 \cdot 10^{-27}$ ergs-seconds.) It

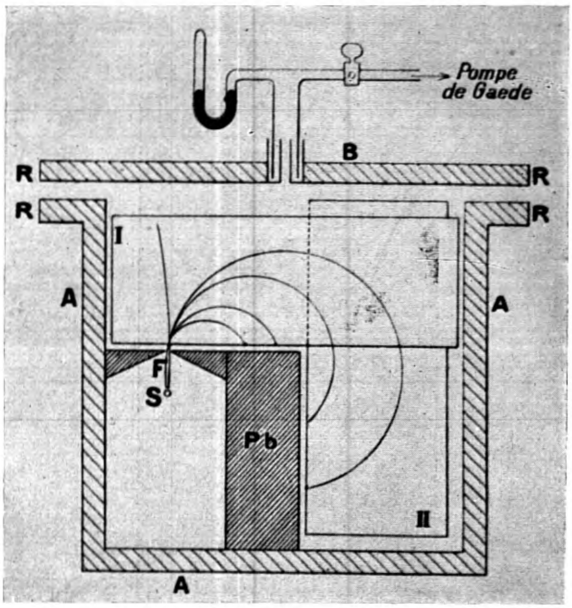


Fig. 1

follows that if an electron is extracted from an atom by this radiation and the work W required to extract it is not exactly as great as the amount $h\nu$, the difference will be turned over to the electron as kinetic energy, and the speed v with which it departs from the atom will be given by the equation

$$\frac{1}{2}mv^2 = h\nu - W$$

and W can be determined by measuring v. We can conveniently refer to W as an "extraction-energy" or "extraction-potential." If all the electrons occupy identical positions, W will be the same for all, and the emerging electrons will all have the same speed. If they occupy various positions or "levels" as is more commonly said, there will be as many different electron-speeds represented in the emerging electron-stream as there are levels,[1] and from these speeds the extraction-energies characterizing (or indeed defining) the levels can be deduced.

The apparatus in which the test is made is of the type shown in Fig. 1. At S there is a long narrow rod or tube of the material being tested, irradiated by X-rays proceeding from a source at the left. A magnetic field, directed normally to the plane of the paper, sweeps

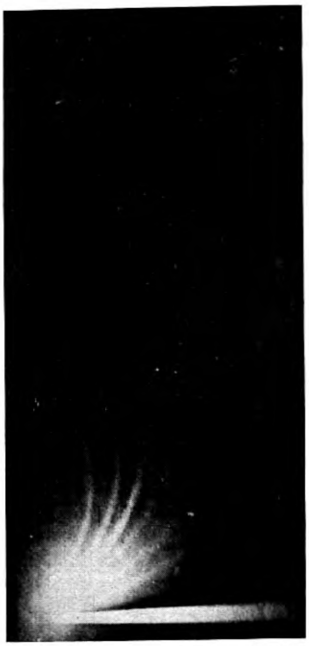


Fig. 2

the emerging electrons around in circular arcs, some of which pass through the slit; a few such arcs are sketched. The slower the electron, the more highly curved the path in which it travels; and the speed of the electron can be deduced from the curvature of the path. In Fig. 2 electron-paths of this type are reproduced from a photographic film, which was laid parallel to the plane of the paper, in the position of the rectangle marked I in Fig. 1. Fig. 3 shows arcs which appeared on a film laid in the position of the rectangle marked II.

[1] Of course there may be reasons why electrons in particular positions cannot often or at all be extracted by radiation, even though there is plenty of energy available.

These distinctly-separated arcs show that the emerging electrons fall into several distinct groups, each characterized by a particular speed. In Fig. 4 we see the traces of the electrons on films laid normally to

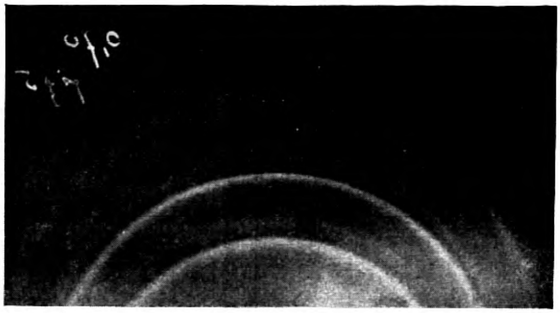


Fig. 3

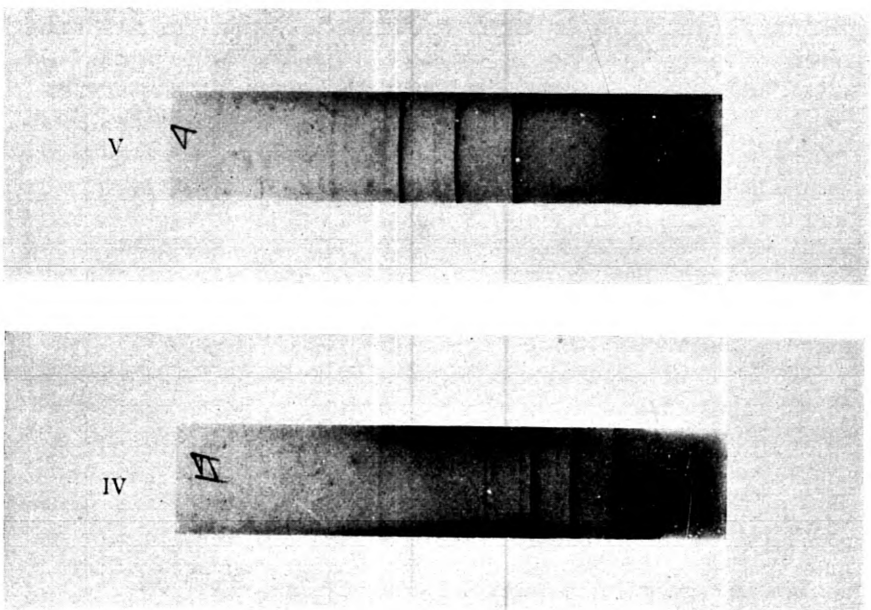


Fig. 4

the plane of the paper, along the top of the block marked "Pb" in Fig. 1. The appearance of the films at once suggests a line-spectrum. The lines, indeed, are the signatures of special electron-speeds instead of special radiation-frequencies; but these two quantities, being interconvertible, are not so profoundly different in their nature as used to be supposed. Imitating de Broglie's term "spectres corpusculaires" we may call these "electronic spectra." But it must be remembered that they depend not only on the properties of the atom, but on the incident radiation as well.

Maurice de Broglie has undertaken an extensive study of these electronic spectra. His most recent apparatus, similar in general to the arrangement illustrated in Fig. 1 (with the photographic plate laid normally to the plane of the arcs) is improved in various respects and enlarged to permit of using a plate 24 cm. wide and electron-paths of 26.cm. radius. Unfortunately, the ideal condition of atoms irradiated by radiation of a single frequency, is unattainable. This is not merely because actual X-ray sources emit very mixed radiations intense at several distinct frequencies and perceptible at every frequency over a wide range. This difficulty could be partly remedied by appropriate filters. There is another difficulty and an inevitable one; the atoms from which electrons are extracted by the radiation promptly emit radiation of new frequencies, which extract other electrons themselves. In the language of the opening paragraph, the arrangement of electrons which results when an electron is extracted is not a permanent one; the remaining electrons redispose themselves in one arrangement after another, eventually arriving at the permanent one; to each successive arrangement corresponds a new and lower value of the energy of the atom, and the energy-differences ΔE are successively sent out in radiations of frequencies $\Delta E/h$. Thus there are several frequencies at work extracting electrons from the atoms; and in the electronic spectrum, each level is represented by as many lines as there are frequencies.

The uppermost spectrum of Fig. 5 is sketched by de Broglie from photographs made with the electrons emitted by silver atoms irradiated with the characteristic X-rays of tungsten.[2] The electron-speeds corresponding to the lines increase from left to right. There are four of these tungsten rays, two forming the $K\alpha$ doublet, while the other two, known as $K\beta$ and $K\delta$, have higher frequencies. The four lines marked 4 and 5 in the electronic spectrum are made by electrons extracted by these four radiations from a single level. This is the K-level, the deepest or innermost level in the silver atom, the electrons removed from it having lost more energy during the removal, than any others observed,—about $3.46 \cdot 10^{-8}$ ergs apiece. The two following doublets, marked 6 and 7, are made by electrons extracted by the $K\alpha$ frequencies from two distinct levels of the silver atom,

[2] Some photographs may be seen in the *Journal de Physique*, volume 2 of 1921. They were taken before the latest improvements were made in the apparatus, and do not show so much detail as the sketches; or perhaps the reproductions are imperfect.

known as the L and M levels respectively; the electrons from them have more energy left over after escaping. Line 8 is due to $K\beta$ extracting electrons from the L-level. The electrons ejected from the M-level by $K\beta$, and those ejected from the L and the M levels by $K\delta$, are presumably moving too rapidly to be received on the plate. At the other end of the spectrum the three lines 1, 2, 3 are due to electrons

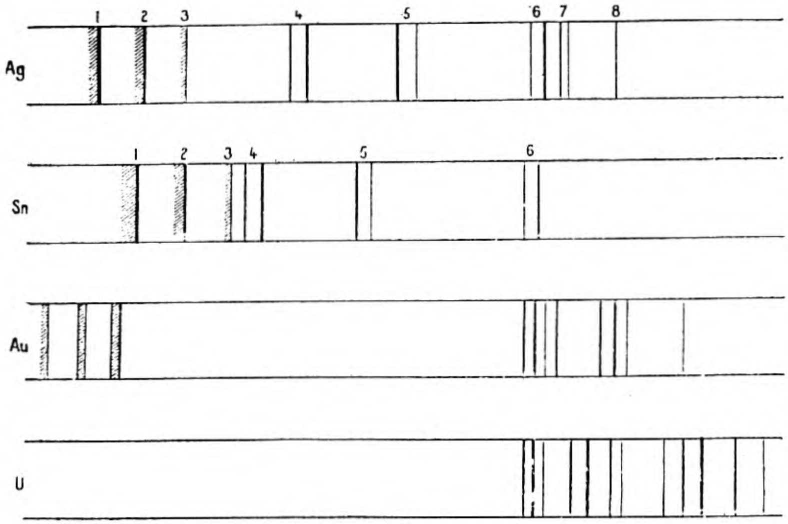


Fig. 5

expelled from the L and the M levels by two of the secondary X-ray frequencies proceeding from silver atoms: the $K\alpha$-doublet (not separated) and the $K\beta$-line of silver.[3] Just below this spectrum, we see the electronic spectrum of tin, in which the lines due to the primary X-rays from the tungsten are arranged like the corresponding lines in the silver spectrum, but displaced towards lower energies, since the levels in the tin atoms are different from those in the silver atom; while the lines due to the secondary X-rays are also repeated from the silver spectrum but with an opposite displacement, for in these cases both the levels from which the electrons are taken and the energies available for taking them out have been changed. Next come the spectra of gold and uranium. Each of these elements has more electrons per atom than the previous two (uranium has more

[3] From the nature of the rearrangements resulting in the $K\alpha$ and $K\beta$ radiations, it follows that the electrons extracted by the former from the M level have the same speeds (very nearly) as those excited by the latter from the L level; the two frequencies acting on the two levels produce three separable lines.

electrons, ninety-two, than any other element). The complexity of the spectra results from this richness of electrons, but the electrons extracted from the L and M levels of gold by its own radiations can be identified.

It is not necessary to provide an X-ray tube to supply the primary radiation; this can be supplied from the nuclei of radioactive atoms mingled with the atoms being tested, or, by examining radioactive substances, we can discover electronic spectra excited by radiations originating at the nuclei of the atoms themselves. Actually these were the earliest electronic spectra discovered; the first to be observed were photographed by von Baeyer, Hahn, and Meitner in 1910, years before the interpretation was made (the frequencies of the nuclear radiations were not then known). The figures 1, 2, 3 and 4, used to illustrate this article, are taken from a paper by J. Danysz, describing work performed in 1911 at the laboratory of Madame Curie in Paris, upon the electrons or beta-rays emerging from atoms of radium B and radium C. The grouping of these electrons, as we now know, results from their being extracted from the various levels by the several nuclear radiations and the inevitable secondary radiations which they produce in their own atoms. The large number of distinct groups (Rutherford and Robinson distinguished sixteen from radium B and forty-eight from radium C) is very likely due to several co-operating causes; there are several frequencies at work, the atoms have large numbers of electrons, and extractions probably occur exceptionally often where the radiations originate so close to the electrons. The earliest electronic spectra produced from non-radioactive atoms were excited by nuclear rays from radioactive substances, and the earliest rule discovered was that these spectra were very similar to the spectra of the radioactive atoms themselves; being indeed identical when the excited atoms are isotopes of the atoms which emit the exciting rays. In a complete account of this topic, many other names would be mentioned, notably those of C. D. Ellis and R. Whiddington.

A recently-published and relatively simple case is that of the radioactive atom, uranium X_1, of which the electronic spectrum is shown in Fig. 6 (from an article by Frl. Meitner). This displays three lines made by electrons of which the speeds indicate that they are extracted from the L, M and N levels of the atom by a single radiation, having itself the frequency of the natural $K\alpha$-radiation of the atom. This radiation was itself detected and identified by appropriate means. Faster electrons which were also observed, cannot have been derived from any such source; they probably came from the nucleus, and some of them eject electrons from the K-level of the atom, thus producing the necessary condition for the $K\alpha$-radiation and all the others to be emitted. These electrons from the K-level would escape with too little energy to be registered in the apparatus. The question of the ultimate origin of these fastest electrons is, however, still under debate by the leading authorities on the subject.

Fig. 6

Imagine now that a beam of X-rays including all frequencies is directed against a thin sheet of metal atoms, and that the transmitted beam is dispersed into a spectrum projected against a photographic plate in the usual manner. Rays of frequency ν can extract electrons from a particular level when $h\nu$ exceeds the value of W for that level, but not otherwise. Advancing along the spectrum in the direction of increasing frequencies, we should expect to find a sudden sharp weakening of the transmitted rays wherever the frequency becomes equal to one of the values W/k which characterize the various levels. Some of L. de Broglie's classical photographs are shown in Fig. 7 (borrowed from Millikan's book, "The Electron"). The second picture from the top represents two spectra of an X-ray beam transmitted through molybdenum, one spectrum stretching away to the right from the central dark band, the other to the left. The frequency decreases as the distance from the dark band increases. Coming inwards toward the band, we see that the plate very suddenly becomes whiter at a certain critical frequency; this is the frequency at which $h\nu$ becomes equal to the W of the K-level. Similar spectra of beams transmitted through cadmium, antimony, barium and mercury are presented below the molybdenum spectrum; the corresponding absorption-edge is discerned in each, its frequency rising with the atomic number of the element.[4] This is by far the most delicate and accurate method of determining the various extraction-energies,

[4] The topmost picture shows the spectrum of the beam before it encounters the absorbing layer; the various strong lines in it, and the absorption-edges impressed upon it by the silver and bromine atoms in the photographic film, recur more or less clearly in the absorption-spectra, but have nothing to do with the atoms in the absorbing layer.

although by itself it merely shows that particular transformations of X-ray energy become possible at particular frequencies. The electronic spectra, although much less accurate for purposes of measurement, are needed to show how this absorbed X-ray energy is used.

These absorption-spectra show that the levels in the atom are much more numerous than the electronic spectra with their lower "resolving power" can reveal; for example, there are three L-levels and five

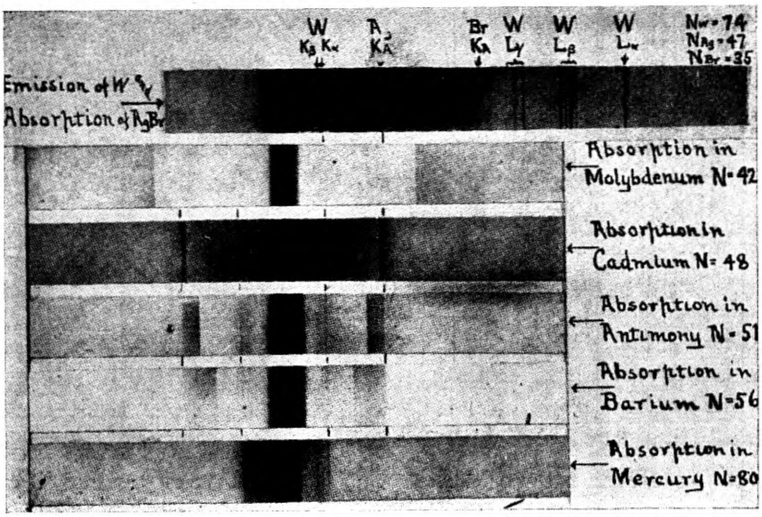


Fig. 7

M-levels. The five M-levels of thorium display themselves (not as clearly as might be desired) in Fig. 8, which consists of absorption-spectra photographed by P. A. Ross at Leland Stanford University. Each spectrum extends from low frequencies at the left to a maximum limiting-frequency at the right, the limit depending on the voltage applied to the X-ray tube and not on the properties of the absorbing atoms. As the limiting-frequency is increased by increasing the voltage, the absorption-edges resulting from electrons being extracted from the five M-levels successively appear. Along with each new absorption-edge there appear one or two new emission-lines, emitted by the thorium atoms during the rearrangements which follow upon the extraction of an electron.[5] These correspond

[5] Actually these lines were not emitted by the same atoms as absorbed the X-rays, but by thorium atoms in the target of the X-ray tube whence the primary X-rays came; the preliminary electron-extractions were performed in most cases by swift electrons. There is no reason to suppose that the agency effecting the electron-extraction has anything to do with the subsequent rearrangements of the atom.

to the secondary radiations of silver and tin, which we found to produce lines of their own in the electronic spectra of these elements. These relations between absorption-edges and emission-lines make

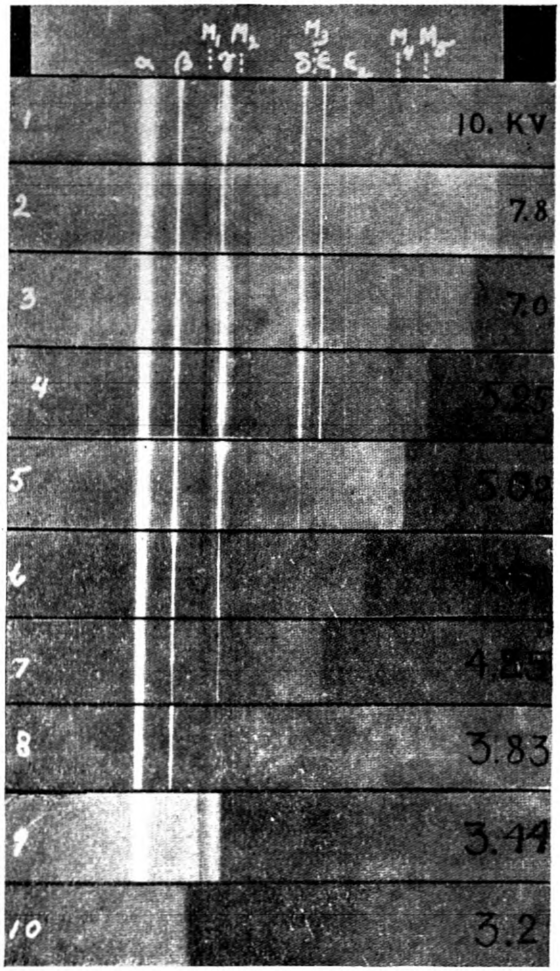


Fig. 8

it possible to use the X-ray emission-lines of atoms to identify and map out their levels.

This discussion of electronic spectra and X-ray absorptions has served to illustrate the remark made in the opening paragraph, that our knowledge about the various arrangements of the electrons forming the atom consists mainly of data about their energy-values. We have a key to the arrangements themselves, and this is provided by the deflections of electrons as they pass through the atoms. An electron shot directly at an atom will be deflected by the combined actions of the nucleus and the atom-electrons; and by postulating a particular arrangement of the electrons we could, in principle at least, calculate the deflection. This may be likened to the performance of an astronomer who, observing a comet advancing into the solar system from outer space, calculates the path which it will follow through the system under the influence of the sun and the major planets, and the direction along which it will depart. The astronomer has the advantages of knowing exactly where the members of the solar system are, and of being able to follow individual comets. We do not know where the members of the electron-system are, and cannot shoot a single electron at an atom and discern its path.

The latter disadvantage is not as serious as it may seem. By projecting an enormous number of electrons in parallel directions against an atom or a layer of atoms, and measuring the fraction which are deviated through a given angle or range of angles, it is possible to test a particular atom-model. Assume that the atom possesses spherical symmetry; then the deflection suffered by an oncoming electron will depend only on a single variable, the minimum distance p from the centre of the atom to the line (extended) along which the electron approaches at first (before the deviation begins). Designating by ϕ the angle between the initial and final directions of motion of the electron (i.e., the amount of the deflection), we have

$$\phi = f(p) \qquad p = f^{-1}(\phi) \tag{1}$$

the function f depending on the particular atom-model. Suppose an enormous number N of electrons directed normally against a thin layer of metal atoms, in which Q atoms lie side by side. The number of electrons which will approach the layer along lines passing some atom-centre—any atom-centre—at distances greater than a given value p and less than a slightly greater given value $p+dp$, is

$$dN = NQ \cdot 2\pi p \cdot dp \tag{2}$$

This is likewise the number of electrons which will be deflected through angles lying between $\phi = f(p)$ and $\phi + d\phi = f(p+dp) = f(p) + (df/dp)dp$; which therefore may be written as

$$dN = NQ \cdot 2\pi p (dp/df) df = F(\phi) d\phi. \tag{4}$$

The expressions for p and dp/df are to be taken from equation (1). The function

$$F(\phi) = NQ \cdot 2\pi p(dp/d\phi) \tag{5}$$

represents the *distribution-in-angle* of the deflected electrons. If it is calculated for any particular atom-model and then determined by experiment, the comparison between calculation and data affords a test of the atom-model. An instructive comparison can be made even if the value of NQ is unknown, since the form of the *F-versus-*ϕ curve, as well as the absolute height of its ordinates, depends upon the atom-model.

For electrons or other charged particles of charge e and mass m, streaming with uniform speed U against a group of much more massive nuclei each bearing a charge E, the functions f and F assume the forms

$$f(p) = 2 \cdot \text{arc cot } (mU^2p/eE) \tag{6}$$

$$F(\phi) = NQ\pi(eE/mU^2)^2 \cot (\tfrac{1}{2}\phi) \operatorname{cosec}^2(\tfrac{1}{2}\phi). \tag{7}$$

This case, insignificant as it may appear, suddenly assumed the greatest importance when, in 1913, Rutherford, Geiger and Marsden established that the distribution-in-angle of alpha-particles (particles of twice the charge and about 7,500 times the mass of an electron) deflected by metal atoms is of precisely the form (7). This means that around each atom-nucleus there is an empty space so wide that full-speed alpha-particles passing close enough to a nucleus to be deflected through 5° or more, undergo almost their entire deflection within it; hence, most or all of the electrons surrounding the nucleus must lie beyond this vacant central region. From the data of these classical experiments, Rutherford and his collaborators deduced that the radius of the empty region encircling the gold nucleus is at least 36×10^{-12} cm. After the war, the problem was again taken up in Rutherford's laboratory in Cambridge. J. Chadwick gave 14×10^{-12} cm. as a minimum value for the radius of the vacant space around the platinum nucleus. Last year P. M. S. Blackett made a statistical study of the deflections of comparatively slow alpha-particles, using the C. T. R. Wilson expansion-method, which was described in the last issue of this Journal. Paths of some of these deflected particles are shown in Fig. 9. By using these slow-moving particles, which begin to turn in their courses while still much farther away from the nucleus than the minimum distance at which fast alpha-particles begin to respond to its repulsion, Blackett was able to search farther out for the outer boundary of the empty space. The deflecting atoms were atoms of argon, each consisting of eighteen electrons surrounding a nucleus; atoms of oxygen with eight electrons apiece, and atoms of nitrogen with seven (the latter two kinds of atoms not being discriminated in the study of the data). Blackett concluded that the

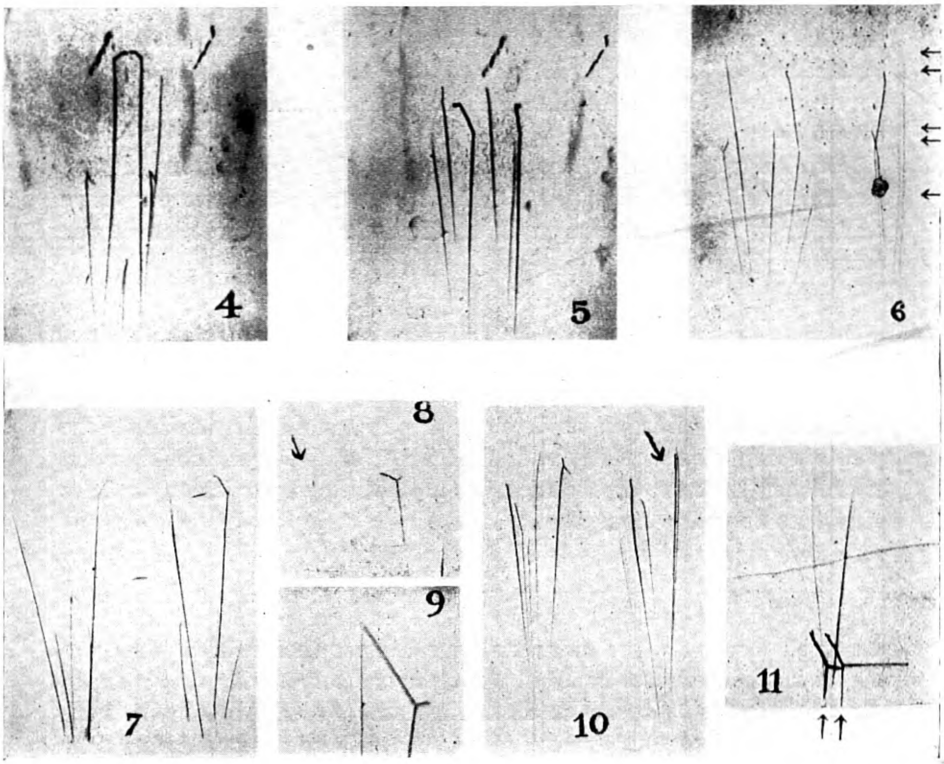


Fig. 9

empty space in the argon atom extends out at least to a distance of 10^{-9} cm. from the nucleus in the argon atom, and to a distance of at least 5×10^{-10} cm. in the nitrogen and oxygen atoms.

We now pass to the case of electrons deflected by atoms. Since the electron is so very much lighter than the alpha-particle, and yet is half as strongly charged, it will be much more seriously deflected by a nucleus than an alpha-particle, approaching along the same line with the same speed, would be. This contrast is very strikingly illustrated by two results published last summer. Harkins and Ryan, photographing the paths of eighty thousand alpha-particles through air, found only three instances of deflections exceeding 90°; C. T. R. Wilson, photographing the paths of 503 fast electrons through air, found forty-four instances of deflections exceeding 90°. While, in general, it would be hardly fair to make such comparisons without allowing for the relative energies of the two kinds of particles, the difference in order of magnitude is so great that we may accept it as typical.

Moreover, the electron will be deflected by the atom-electrons as well as by the nucleus, and will not disarrange the atom-electrons so badly on its way through the atom-system. These deflections

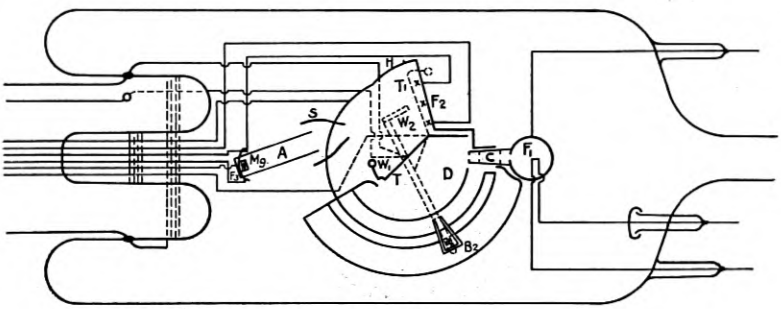


Fig. 10

will be superposed upon the deflection produced by the nucleus, and will modify the distribution-in-angle function F from form (7) into some other form. Such modifications have been suspected by several investigators; for example, by Crowther and Schonland in their study of the deflections of very fast electrons by metal atoms. It has been argued by Wentzel, however, that the distribution-in-angle function observed in their experiments departed from the form (7) not because the atom-electrons were interfering with the fast electrons, but because some of the deflected electrons had been deviated by several atom-nuclei in succession.

C. Davisson and C. H. Kunsman, in the laboratories of the Western Electric Company, made the first definite attempt to produce electron-deflections under conditions in which the distribution-in-angle function would disclose the influence of the atom-electrons. To do this it was desirable to use, not the fast electrons from radioactive atoms which previous experimenters had employed, but slow electrons of controllable speed. A diagram of their apparatus is shown in Fig. 10 and a photograph in Fig. 11. The electrons proceed from a hot filament at F_1, strike the metal target at T, and are deflected through various angles; the shielded collector B_1, swinging from one angle to another, successively receives the electrons deflected through the various angles. The electrons depart from F_1 with very low speeds and receive the speed U through acceleration by a voltage applied between F_1 and the cylinder surrounding F_1; thereafter they move with constant speed in an equipotential region, through the various

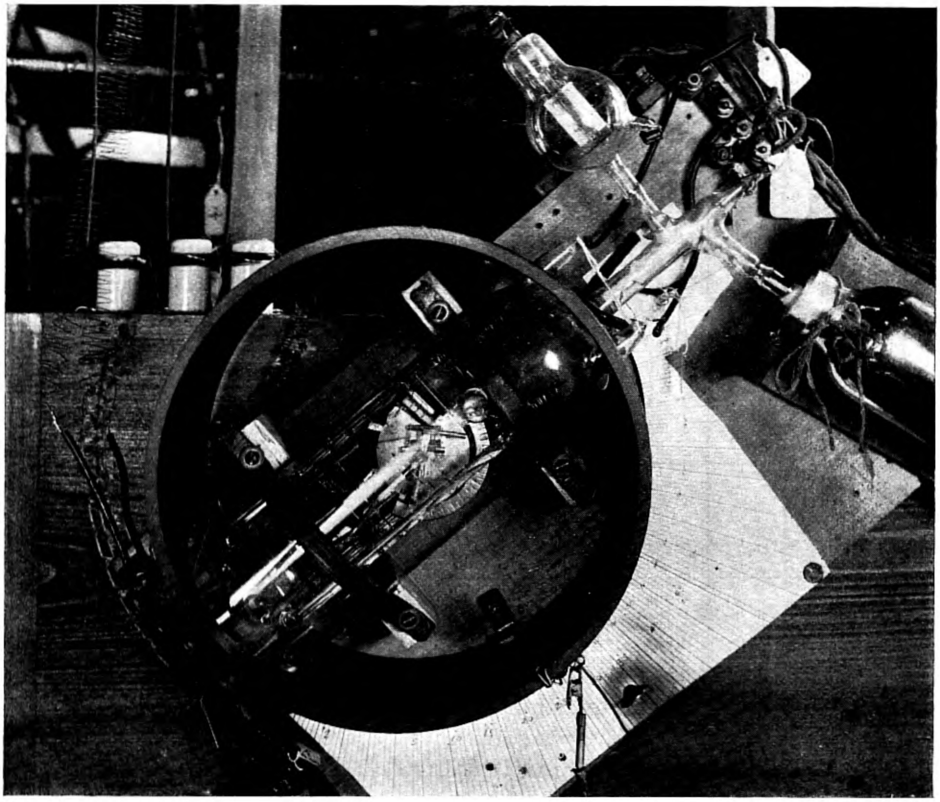


Fig. 11

slits shown in the diagram around C, and against the target. Among the electrons which emerge from the target, there are some which have been deflected by individual atoms in the way we have been describing, but very many more which have either undergone several deflections in succession or else were not in the incident beam but have been dislodged from their places in the target metal by the primary electrons. If these latter were allowed to reach the collector, the distribution-in-angle function of the once-deflected electrons would be blurred and concealed by the unwanted electrons. As, however, they have not so much energy as the primary or the once-deflected electrons, they can be kept away from the collector by lowering its potential to a value such that only such electrons as have, say, 90% of the energy of the primary electrons can reach it. Thus the filament may be at potential zero, the target at 500 volts; if the collector is also at 500 volts, the distribution-in-angle function of the

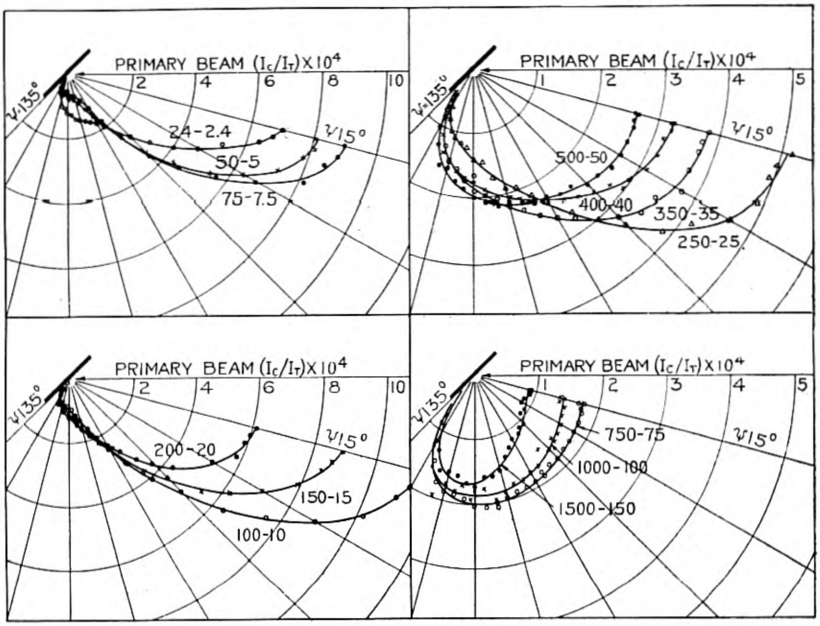


Fig. 12

electrons it receives has nothing in common with the function F characterizing the once-deflected electrons; but when the collector is lowered to 50 volts, the distribution-in-angle function which it records assumes a new and characteristic form.

Some of these angular distributions are shown in Fig. 12 (for magnesium) and Fig. 13 (for platinum). The latter curves were obtained first, with a platinum target; then the target was overlaid with a thin film of magnesium, formed by sublimation without opening or altering the tube, and the sharply-contrasted curves of Fig. 12 replaced the others. The distribution-in-angle of the engineering electrons is plotted, naturally, in polar coordinates; the direction $\phi = 0°$, i.e., the direction of motion of the primary electrons, is indicated by the arrow and the lettering. Such a symbol as "100–10" indicates that the corresponding curve was taken down with the target at 100 volts and the collector at 10 volts (the filament always being at zero potential). The reason for this has been explained above; the family of curves in Fig. 12 illustrates the point.

These are examples of the curves from which the arrangement of the atom-electrons is to be inferred. The sinuous and serrated curves for platinum, entirely different from the smoothly rounded curves

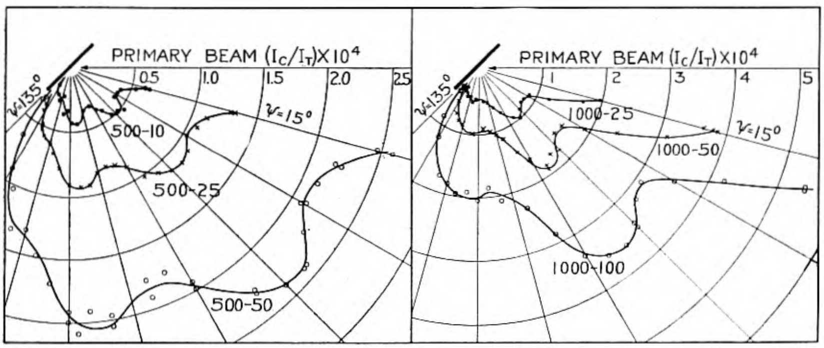


Fig. 13

derived from equation (7), surely owe their shape to the numerous levels among which, as was shown in the foregoing pages, the electrons of massive and electron-rich atoms are distributed; the platinum atom, with its seventy-eight electrons, ranks among the most complicated of all. The magnesium atoms, with their thirteen electrons apiece, are simpler and yield curves which are simpler, but not of the type of equation (7).

To interpret these curves Davisson has calculated the distribution-in-angle function for electrons deflected by an idealized "limited-field" atom-model, in which there is a concentrated charge $+E$ at the centre and a charge $-E$ uniformly spread over a spherical surface of radius R. This uniformly-charged sphere is a sort of first-approximation substitute for a spherical surface on which several electrons are arranged. It is not implied that the magnesium atom has all its electrons at the same distance from the nucleus, which would be most improbable, as its X-ray spectrum shows at least two distinct levels; we can suppose that n out of the 12 electrons are so close to the nucleus that together with it, they practically form a single point-charge $(12-n)e$, and the remaining $(12-n)$ electrons lie on the spherical surface, which encloses the "empty space" mentioned above. The functions f and F assume the forms

$$f(p)=2 \text{ arc } \cot\frac{p(2\mu-1)}{\sqrt{R^2-p^2}} \tag{6a}$$

$$F(\varphi)=NQ\pi \cot\left(\tfrac{1}{2}\phi\right) \operatorname{cosec}^2\left(\tfrac{1}{2}\phi\right) \frac{(2\mu-1)^2R^2}{[\cot^2(\frac{1}{2}\phi)+(2\mu-1)^2]^2}, \tag{7a}$$

in which $\mu=\frac{1}{2}mU^2R/eE$; the symbols have the same meanings as in (6), and (7), with which these equations become identical if R is made infinite.

This "limited-field" distribution-function has some odd characteristics. At very high speeds large deflections naturally are rare, but as the speed is lowered they become relatively more frequent; the 1,500-volt, 1,000-volt and 750-volt curves for magnesium illustrate this. This tendency gains rapidly as U is decreased; at a certain critical value, given by $\mu=1$, the deflections are uniformly distributed in all directions[6]; at a lower critical value, given by $\mu=\frac{1}{2}$, all the electrons are turned through 180° and return on their tracks. As the speed is still further decreased, the condition of uniformly-distributed deflections is again approached, and we have the extraordinary feature of the average deflection decreasing as the energy of the electrons goes down.[7] In the family of curves for magnesium there appears very clearly an intermediate velocity at which 180° deflections are peculiarly frequent; the curves spread outward in the direction $\phi=180°$ whence the primary electrons come, as the energy of the primaries rises from 24 to 75 volts, and retract themselves again as the energy rises beyond 100 volts. This is a particularly important feature of the curves.

To make an adequate test of the new expression for F, it is necessary to apply certain corrections to the curves presented, particularly a correction required because the distance travelled by the deflected electrons within the target metal varies with ϕ, so that the percentage which goes astray, owing to loss of speed or otherwise, varies similarly. A curve exempt from this correction can, however, be ob-

[6] Meaning that the number deflected per unit solid angle is independent of ϕ, which means that the distribution-in-angle function is of the form *const.* (*sin* ϕ).

[7] It may be recalled from the last number of this Journal, page 110, that H. A. Wilson used this property as an explanation of the anomalous variations of electron-mean-free-paths with speed in various gases.

tained in a certain manner.[8] On studying this curve, it is found that the critical speed at which 180° deflections are most frequent is too low. This indicates that an incident electron approaching an atom is accelerated toward it, by virtue of the total charge of the electrons on the spherical shell not quite compensating the nuclear charge; the speed U which figures in the equations is therefore greater than the measured speed with which the electrons are fired at the target. (This interpretation also serves to explain the lobe observed on the lowest-speed curves for magnesium, and suggests the reason for the lobes of the curves for platinum.)

The curves are satisfactorily explained, if we build the magnesium atom in this manner: a nucleus of charge $12e$, two electrons so near it that the central charge is effectually $10e$, and a spherical shell of six electrons with a radius of $1.28 \cdot 10^{-9}$ cm.; the other four electrons much further out, perhaps dispersed and wandering through the metal. The only arbitrary assumption made is that about the two deep-seated electrons; the radius R of the shell and the number of electrons upon it are prescribed by the curves, once that assumption is made. If we assume three deep-seated electrons, R becomes $1.15 \cdot 10^{-9}$ cm. and the number of electrons in the shell drops to five. The shell must be the L-level, and the deep-seated electrons constitute the K-level.[9]

The energy required to remove the loosest or outermost electrons of the atom is generally determined, as is well enough known, by smiting the atom with an electron instead of with one of the radiation quanta used in extracting the inner electrons.[10] Usually the quantity measured is simply the energy which the striking electron must have, in order to convert the atom or molecule into a positively-charged ion; the negative charge removed from the atom is assumed without proof to be a single electron. On the other hand, J. J. Thomson

[8] Imagine an electron incident at angle θ on the target surface, and deflected through angle ϕ (in the plane of incidence) by an atom which it meets after penetrating a distance d in a straight line. If it continues in a straight line from the point of deflection until it emerges, it travels a distance $x = d\,(1+\cos\theta \cdot \sec(\psi-\theta))$, where $\psi = \pi - \theta$. This distance x will be the same for any two values ψ_1 and ψ_2 of ψ, such that $\psi_1 + \psi_2 = 2\theta$. Insofar as the number of deflected electrons emerging with speed sufficient to reach the collector depends on x, it will be the same for both values of ψ. The curve representing the ratio of the number of electrons reaching the collector, for two such angles, plotted versus U, is exempt from this correction, and can be directly compared with a theoretical curve.

[9] Or we could assume that there were no deep-seated electrons, and give seven electrons and a radius $1.54 \cdot 10^{-9}$ cm. to the shell; but then we should have nothing to serve as a K-level.

[10] Generally the frequency required to extract the outermost electron with a quantum lies in the most inconvenient region of the spectrum for practical work.

and many others have measured the charges [11] of ionized atoms in discharge-tubes, and found them sometimes single and sometimes multiple electron-charges, but have not measured the minimum energy required to produce a particular kind of ionization. H. D. Smyth, at Princeton and Cavendish, was the first to combine both methods; he ionized atoms by electron-impacts in a tube designed for determining ionization-potentials in the accepted manner, and after further accelerating the ions drew them through a channel into a second tube where they were deflected in a magnetic field so that their charges could be measured. The difficulty to be surmounted is that in the first tube the pressure of the gas must be high enough to yield a satisfactory number of ions, and in the second tube it must be low enough not to interfere with the arcs described by the ions in the magnetic field. At first he sent a beam of mercury vapor rushing transversely across his first tube from a boiler into a liquid-air trap; by first sending the atoms down a long tube with a system of diaphragms and so stopping the obliquely-moving ones, he was able to prevent atoms from straying out of the beam in the critical zone. Later he attacked a more difficult case, that of nitrogen; the gas was continuously fed into the first tube and a powerful pump drew it out before it could diffuse seriously into the second tube.

While it is interesting to have direct confirmation that the first and easiest ionization is the extraction of a single electron, Smyth's most important results refer to the later ionizations. Mercury atoms that had lost two electrons appeared in the second tube when the bombarding potential attained 19 volts, nine volts more than the first ionizing-potential; at a much higher voltage, triply-charged atoms were detected, or at least suspected. In nitrogen, the earliest ionization, at about 16 volts, does not involve dissociation, but at a potential 8 volts higher, a doubly-ionized single nitrogen atom makes its appearance, and a little further along, Smyth detects an ion which may be a singly-ionized nitrogen atom or a doubly ionized molecule (the two possibilities cannot be discriminated by this method, but the second seems improbable). Valuable knowledge about the relations between ionization and dissociation—between, that is, the removal of an electron from a molecule, and the breaking of the bonds that hold the atoms of the molecule together—may be expected from experiments of this type.

Something more is to be said on two of the topics of the last article in this series. A. H. Compton's discovery that scattered X-rays consist of two distinct radiations, one with the frequency of the

[11] Actually, the charge-mass ratios.

primary rays and the other with a slightly lower frequency, was mentioned in that article; he has since published an account of a series of measurements made, not on the wave-length but on the absorption-coefficient (in various substances) of the scattered rays, and finds it altered from that of the primary rays in the sense and more or less in the magnitude to be expected from the wave-length measurements of the lower-frequency rays. The largest alterations and the best agreements with theory are obtained with light atoms and high-frequency rays. In the frequency-range of the visible spectrum, the scattered ray of lowered frequency, sought for by P. A. Ross in light of the wave-length 5461A scattered by mercury vapor, is altogether lacking. The transparency of krypton and xenon atoms to slow electrons, discovered by Minkowski and Sponer, has been confirmed by Ramsauer with his original (and better) method. The transparency of argon atoms has also been verified by O. W. Richardson and R. N. Chaudhuri, by a method sufficiently different from the others to rank as an independent test.

REFERENCES

O. von Baeyer, O. Hahn, and L. Meitner: Phys. ZS, *11*, pp. 488-493 (1910); *12*, pp. 273-279 (1911).

P. M. S. Blackett: Proc. Roy. Soc. *102A*, pp. 1-17 (1922).

L. de Broglie: Jour. de Phys. (5) *6*, pp. 161-168 (1916).

M. de Broglie: Jour. de Phys. (6) *2*, pp. 265-267 (1921); C. R. *173*, *172* (1922).

J. Chadwick: Phil. Mag. *40*, pp. 734-746 (1920).

A. H. Compton: Phil. Mag. *46*, pp. 897-911 (1923).

J. Danysz: Le Radium *9*, pp. 1-6 (1912); *10*, pp. 4-6 (1913).

C. Davisson, C. H. Kunsman: Science, November 25, 1921 and later issues; Phys. Rev. *21*, pp. 637-649 (1923); Phys. Rev. *22*, pp. 242-258 (1923).

C. D. Ellis: Proc. Roy. Soc. *101A*, pp. 1-17 (1922).

H. Geiger and E. Marsden: Phil. Mag. *25*, pp. 604-623 (1913).

W. D. Harkins and R. D. Ryan: Jour. Am. Chem. Soc. *45*, pp. 2095-2107 (1923).

L. Meitner: ZS. f. Phys. *17*, pp. 54-66 (1923).

C. Ramsauer: Ann. d. Phys. *72*, pp. 345-352 (1923).

O. W. Richardson, R. N. Chaudhuri: Phil. Mag. *45*, pp. 337-352, and *46*, pp. 461-472, 553-564 (1923).

P. A. Ross: Phys. Rev. *22*, pp. 201-202 (1923). (Scattering of light.)

P. A. Ross: Phys. Rev. *22*, pp. 221-225 (1923). (X-ray spectra.)

E. Rutherford: Phil. Mag. *21*, pp. 669-688 (1911). (Deflection of charged particles by atom-nuclei.)

E. Rutherford and H. Robinson: Phil. Mag. *26*, pp. 717-729 (1913).

H. D. Smyth: Proc. Roy. Soc. *102A*, pp. 283-293 (1922-23); *104A*, pp. 121-134 (1923).

G. Wentzel: Phys. ZS, *23*, pp. 435-436 (1922). Ann. d. Phys. *69* (1922).

R. Whiddington: Phil. Mag. *43*, pp. 1116-1126 (1922).

C. T. R. Wilson: Proc. Roy. Soc. *104A*, pp. 192-212; 1923.

Contributors to This Issue

S. P. Shackleton, B.S., E.E., University of Michigan, 1915. American Telephone and Telegraph Company, Engineering Department, 1915–19; Department of Development and Research, 1919—. Mr. Shackleton's work has been connected with toll circuit development, including toll switchboards, telephone repeaters and testboards.

H. W. Purcell, B.S., Harvard University, 1905; Engineering Department of the Western Electric Company, 1906–20; Department of Development and Research, American Telephone and Telegraph Company, 1920—. Mr. Purcell has been associated with the design of electromagnets and more recently with the development of machine switching apparatus.

Walter A. Shewhart, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company, 1918—. Mr. Shewhart has been engaged in the study of the relationship between the microphonic and the physico-chemical properties of carbon.

Edward C. Molina, Engineering Department of the American Telephone and Telegraph Company, 1901–19, as engineering assistant; transferred to the Circuits Design Department to work on machine switching systems, 1905; Department of Development and Research, 1919—. Mr. Molina has been closely associated with the application of the mathematical theory of probabilities to trunking problems and has taken out several important patents relating to machine switching.

Robbins P. Crowell, Southern New England Company, assistant chief operator, New Haven, 1907–09; chief operator, New London, 1909–10; New York Company, assistant manager, 1910–13; engineer in the Traffic Department, 1913–16; Engineering Department, American Telephone and Telegraph Company, 1916—. Mr. Crowell has been concerned chiefly in the preparation and introduction of peg counts and force adjustment routines.

Francis F. Lucas, Plant and Engineering Departments of Associate Companies, 1902–10; Installation Department Western Electric Company, 1910–13; Engineering Department, 1913—. Mr. Lucas is a materials engineer and has specialized in the application of microscopy to industrial problems.

J. G. Ferguson, B.S., University of California, 1915; M.S., 1916; research assistant in physics, 1915–16; Engineering Department of the Western Electric Company, 1916—.

Karl K. Darrow, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917—. At the Western Electric, Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

ERRATA

ISSUE OF JANUARY, 1924

On page 162, line 11 from bottom of page, and on page 163, line 4:

read $K\gamma$ *instead of* $K\delta$.

On page 173, line 3 from bottom of page:

read emerging *instead of* engineering.