In the days following its historic landing on Comet 67P/TG on 12 November 2014, the Philae module – dropped by Europe's Rosetta probe[1] – sent numerous images and data that will ultimately help scientists to better understand the genesis of the solar system.

Christian Genest and James Hanley*** [2] *use the planning of this 10-year mission to* ***connect several statistical distributions****.* [3]

This scientific feat, achieved 500 million kilometers from Earth, was the culmination of a perilous journey undertaken by the Rosetta probe on 2 March 2004. The mission, which took 10 years of preparation and investments valued at over 1 billion euros, also represented a major technical challenge.

In its journey of more than 6.5 billion kilometers, which lasted 10 years, the probe had to withstand the thermal and energy plans large amplitude variations of solar lighting imposed by its trajectory. Electronic components have also been exposed to high doses of radiation due to solar flares.

The question of the reliability of equipment and optimization of relief materials arises in any mission where re-stocking is not possible. It is even more crucial in cosmic journeys, subject to strong constraints of weight and space. We must avoid unnecessary duplicate systems but if we 'go too close to the edge,' we risk failure.

Since failures are unpredictable - and therefore stochastic - optimizing the number of spare parts to be carried is based on probability calculations. While taking some poetic licence, we illustrate some of these, using the ambitious Rosetta project of the European Space Agency.

If resources are limited

Imagine that a certain piece of electronic equipment essential to the proper functioning of Rosetta wears out so slowly that, in practice, only the plasma stream of solar wind can affect it. When this piece fails, it is immediately replaced by an identical piece, and so on, while stocks last. If the mission carries k identical pieces, the system can operate continuously until the occurrence of the kth failure.

Specifically, suppose a burst of solar wind shakes the probe on average once every 16 months (4/3 years), at random (unpredictable) times that are independent of each other. Thus, one would expect to have an average of $\lambda = 3/4$ failures per year, or $\lambda \times 10 = 7.5$ failures in 10 years. If the probe carries only $k = 3$ pieces, then there would be little chance that it is still operational at the rendezvous with comet 67P/TG ("Tchouri" for short).

More generally, what are Rosetta's chances of success if it carried 6, 9, 12 or 18 identical pieces? To quantify the risk as a function of the total number of pieces carried, we must make assumptions, about either the distribution of the number of failures, or the distribution of the life of each piece. This is where the theory of probability and statistics comes to the rescue of the decision maker.

Two complementary stochastic models

Let N be the unknown number of failures that occur during a mission of t years. It is reasonable to assume that this variable (roughly) follows a Poisson distribution, named after the French mathematician Siméon Denis Poisson (1781-1840) who proposed it. This model states that if $\mu_N = \lambda t$ is the average

[1] See `http://www.esa.int/Our_Activities/Space_Science/Rosetta` for more details on this mission. Under the headline "Mission complete: Rosetta's journey ends in daring descent to comet" the website announced: 'On 30 September 2016 ESA's historic Rosetta mission has concluded as planned, with the controlled impact onto the comet it had been investigating for more than two years.'

[2] * Mathematics & Statistics, ** Epidemiology, Biostatistics & Occupational Health, McGill University

[3] This piece is adapted from the Acromath article by Genest and Hanley - see Resources

number of outages during a mission of t years (i.e. λ per *year* of operation), then for each $n \in \{0, 1, \dots\}$, the probability (Pr) that $N = n$, i.e., the probability that one observes n failures during t full years of operation, is given by

$$\Pr(N = n) = \frac{\mu_N^n \; e^{-\mu_N}}{n!}$$

where $n! = 1 \times 2 \times \cdots \times n$ for $n \in \{1, 2, \dots\}$ and $0! = 1$ by convention.

If the mission carries k identical pieces, a t-year mission will be completed successfully if at most $k - 1$ pieces fail. The probability of this event is $\Pr(N \leq k - 1) = \Pr(N = 0) + \Pr(N = 1) + \;\ldots\; + \Pr(N = k - 1)$. If $\lambda = 3/4$ and $t = 10$, and if $k = 12$, this probability is approximately 0.921 (or 92.1%), as can be calculated with the Excel spreadsheet using the command POISSON (11; 7.5 ; cumulative = TRUE).

Another, equivalent, way to proceed would be to consider the total life of the system. If the module is equipped with k pieces and L_i denotes the lifetime of the i-th piece, then the total lifetime is given by

$$L = L_1 + ... + L_k$$

and the mission will be a success if $L > t$ years. The identity $\Pr(L > t) = \Pr(N \leq k - 1)$ represents a perfect link between the discrete distribution of the number N of failures and the continuous distribution of the lifetime, L, of the system.

Since the pieces are identical, it is reasonable to assume that their lives $L_1, \dots, L_k$ are mutually independent and identically distributed. If the number N of failures in t years of operation is a Poisson variable whose average value is λt, we then have

$$Pr(L_i > t) = e^{-\lambda t},$$

for all all $i \in \{1, ..., k\}$ and $t > 0$. In effect, with $0! = 1$, we see that

$$Pr(L_i > t) = \Pr(L_1 > t) = Pr(N = 0) = e^{-\lambda t}.$$

We say that the variables $L_1, \dots, L_k$ are exponentially distributed random variables, each with failure rate λ/year. Thus the average lifespan of each piece is $(1/\lambda)$ years. Since they are k in number, the life expectancy of the system is k/λ years of operation. This is good to know, but what interests especially the mission planners is $\Pr(L > t)$.

The Danish mathematician Agner Krarup Erlang (1878-1929), who was very interested in the theory of queues and management of telephone systems, determined the probability distribution of the sum $L = L_1 + \cdots + L_k$ of k mutually independent exponentially distributed random variables each having a rate parameter λ. He showed that if one could observe an infinite number of values of L, the shape the histogram would be given by the formula

$$f_L(l) = \frac{\lambda(\lambda l)^{k-1} \; e^{-\lambda l}}{(k-1)!}.$$

The probability that $L > t$ is then given by the area under the curve between the points t and infinity, namely

$$\Pr(L > t) = \int_t^\infty f(l) dl.$$

The law of Erlang being a special case of the gamma distribution, the complement of this integral can be evaluated with the command of Excel:

$$GAMMADIST(t; 12; 4/3; cumulative = TRUE)$$

When we put $t = 10$, the Excel result is 0.079; its complement is $1 - 0.079 = 0.921$, or 92.1%.

The link between the discrete and continuous

Performing the change of variable $w = \lambda l$ in the above integral, and taking into account that $dw = \lambda\ dl$, we find that

$$\Pr(L > t) = \int_t^{\infty} \frac{w^{k-1}e^{-w}}{(k-1)!} dw.$$

Since $\Pr(L > t) = Pr(N \leq k-1)$, we deduce that

$$\sum_0^{k-1} \frac{\mu^n e^{-\mu}}{n!} = \int_t^{\infty} \frac{w^{k-1}e^{-w}}{(k-1)!} dw.$$

This remarkable identity, valid for any integer k in $\{1, 2, ...\}$, can also be demonstrated by recursion (see Appendix 1). It establishes a direct and precise link between the upper tail area of the (continuous) gamma distribution and a finite sum representing the corresponding lower tail of the (discrete) Poisson distribution. Thus, one can evaluate one or the other, according to the calculation means at hand.

In the early 20th century, there were no electronic computers let alone Excel spreadsheets. In his seminal 1900 work on a statistical goodness-of-fit test for frequency data, the English mathematician Karl Pearson (1857-1936) showed several examples, and for each one calculated the *upper* tail area of (what we now know as) the chi-squared distribution by computing a finite sum. If one examines the sum carefully (**see example**) one will immediately recognize that this upper tail is also a sum of Poisson probabilities for the integers 0 to some $(k-1)$. His 1900 paper provided tables for df. With time, tables of values of Erlang integral were constructed for various values of k and μ.

Another founding father of modern statistics, the Englishman Sir Ronald Fisher (1890-1962), often exploited the identity shown in the box to avoid tedious calculations, and his 1934 article connected many of the common distributions.

Even today, Excel and many other statistical software implicitly make use of this identity. The current methods of numerical calculation allow integrals to be evaluated with great precision, even when the rounding errors involved in the evaluation of the different terms of a sum are cumulative; if the sum includes many terms, the accumulated error can be significant even if each of them is very small, so care is needed.

Other applications and generalizations

The Poisson-exponential model presented here has many applications, not only in aerospace and theory of queues, but in many areas, such as insurance, biochemistry or the study of disease. The requirements for applying the model are those of a “Poisson process.” Even though they are relatively few contra-indications, these are not always checked in practice. One may observe an abnormally high number of zeros. Or, it may happen that the life spans of successive parts are neither exponentially distributed nor independent of each other. In our example, a particularly violent solar storm would likely damage all Rosetta spare parts simultaneously. In such contexts, the probability calculations presented here are no longer valid. Various generalizations and variants of the Poisson-exponential model are used to account for such situations.

A telling figure

All the concepts presented in the article can be illustrated nicely by the figure below, produced using the statistical freeware R [see Appendix 2]. The figure is an assembly of five panels labeled A, A’, B, C and D. We now explain what each one represents.

Panel A contains 25 lines composed of coloured segments. Every line represents a possible realization of the journey undertaken by the Rosetta spacecraft over a period of 10 years. The lengths of the coloured segments show the successive lives of the pieces, assuming that initially the probe has on board 12 pieces (numbered 1 to 12), and that pieces have mutually independent lives and exponential distributions with $\lambda = 3/4$ failures per year. The number that appears at the end of the line indicates which piece is operating when, after 10 years, Rosetta released the Philae module that landed on the Tchouri comet.

In this simulation with R, the mission proved successful in 22 of 25 cases, or in 88% of the simulated instances. Missions 8, 18 and 24 (counting from the bottom) ended in failure. This proportion of observed successes is quite close to the 92.1% predicted by the theory.

Panel B shows, on the same time scale as panel A, when each of the 25 simulated Rosetta missions would eventually fail for lack of spare pieces. The first three points are those that lead to the failure of simulated missions 8, 18 and 24. In the best scenario, Rosetta would become dysfunctional at the end of 24 years.

While 25 simulated scenarios were enough to get a reasonably acceptable estimate of the probability of failure for a mission of 10 years, the dispersion of the points along the right side of panel B gives only a rough idea of the distribution of the variable L, the total life of the system. Obviously, it is more difficult to estimate an entire curve than to estimate a number! To better understand the distribution of L, we would need to have thousands of repetitions and then to smooth the histogram of the values we obtained. This is often the approach used by the statisticians in very complex situations, involving many variables. In our case, as we have already explained, an exact calculation of the L distribution was possible.

The idealized histogram (the $f(l)$ of the Erlang law) is drawn in panel C for k = 3, 6, 9, 12 and 18 pieces, when the rate is $\lambda = 3/4$ failures per year. The curve corresponding to the case of $k = 12$ pieces is shaded. The dark grey part, which corresponds to cases where $l > 10$ years, represents 92.1% of the area under the curve; the complement, 7.9%, is the theoretical probability that the mission fails. Obviously if we were limited to $k = 3$ or 6 pieces, the chances of success would be small (2% and 24.1%). In contrast, some 99.9% of missions would be successful, if the mission carried $k = 18$ pieces. Thus, the calculation of probabilities enables us to quantify the risk and find a good compromise given the constraints of weight, space as well as cost.

Panel D illustrates the relationship $\Pr(L > t) = Pr(N \leq k - 1)$. This can be read in two complementary ways. The x-axis runs from 0 to 10 years; the ordinate represents probability, expressed in percentages. The colour ranges identified as 1 to 12 are associated with the lives of the parts; the visible white part in the upper right corner represents a failure of the mission (the colour code is the same as in panel A).

To read panel D, let us establish a vertical line, say at time $t = 1$ year. The length of the green vertical segment then represents the probability that piece number 1 is still operational at time t. The length of the next segment is the probability that piece number 2 is the one in operation at that moment, and so on for the other segments. If we now slide t to the right, to say $t = 2$ years, we see that the green segment is smaller. This corresponds to the fact that as time passes, the chances of survival of the first piece are dwindling; in fact, they decrease at an exponential rate, as we have seen previously.

Similarly, the upper boundary of zone $i \in \{1, ..., 12\}$ represents the survival function of the variable $L_1 + \cdots + L_i$, that is to say the curve $Pr(L_1 + \cdots + L_i > t)$ drawn as a function of t. As we have seen, this probability is also the probability of observing $N \leq i - 1$ failures in the time interval $[0, t]$. When $i = 12$, the probability represents the probability that the mission is successful up to then (i.e., up to t); as might be expected, it decreases over time and is 92.1% after 10 years.

Panels C and D are based on theoretical calculations, while panels A and B are the result of a simulation, repeated 25 times. If the assumptions underlying the theory are valid, the resulting formulas predict the portrait that would be realized if one had carried out a very large number of independent trials. Thus, for example, the distribution of colours along the vertical line at $t = 10$ years give the chances that a particular piece is in use at the time the probe reaches the comet. Meanwhile, panel A' shows the empirically observed distribution (we simply reordered the results already reported on panel A). Given the small number of repetitions, the correspondence is quite good!

Box 1: Demonstration of the Identity

The proof is by induction on the number of pieces, k. Note first that the identity is true when $k = 1$ since $0! = 1$, so that

$$\int_{\mu}^{\infty} \frac{w^0 e^{-w}}{0!} dw = e^{-\mu} = Pr(N = 0) = Pr(N \leq 0).$$

Suppose then that the relation holds for any integer k. To show that it remains true for $k + 1$, you only have to integrate by parts. Knowing that $-e^{-w}$ is the derivative of e^{-w} and that $\frac{d}{dw}(w^k) = kw^{k-1}$, one has

$$\int_{\mu}^{\infty} \frac{w^k e^{-w}}{k!} dw = \frac{\mu^k e^{-\mu}}{k!} + \int_{\mu}^{\infty} \frac{kw^(k-1)e^{-w}}{k!} dw.$$

The first term of the sum is $\Pr(N = k)$ and the second is $\Pr(N \leq k - 1)$ by the induction hypothesis. The integral is therefore equal to

$$\Pr(N = k) + \Pr(N \leq k - 1) = \Pr(N \leq k),$$

as stated. This proves that the result is valid for any integer $k \in \{1, 2, \ldots\}$.

Box 2: The freeware `R`

`R` is a programming language and software environment for statistical computing and data analysis. With excellent graphics capabilities in 2 and 3 dimensions, this software is royalty free and may be downloaded from the following site:

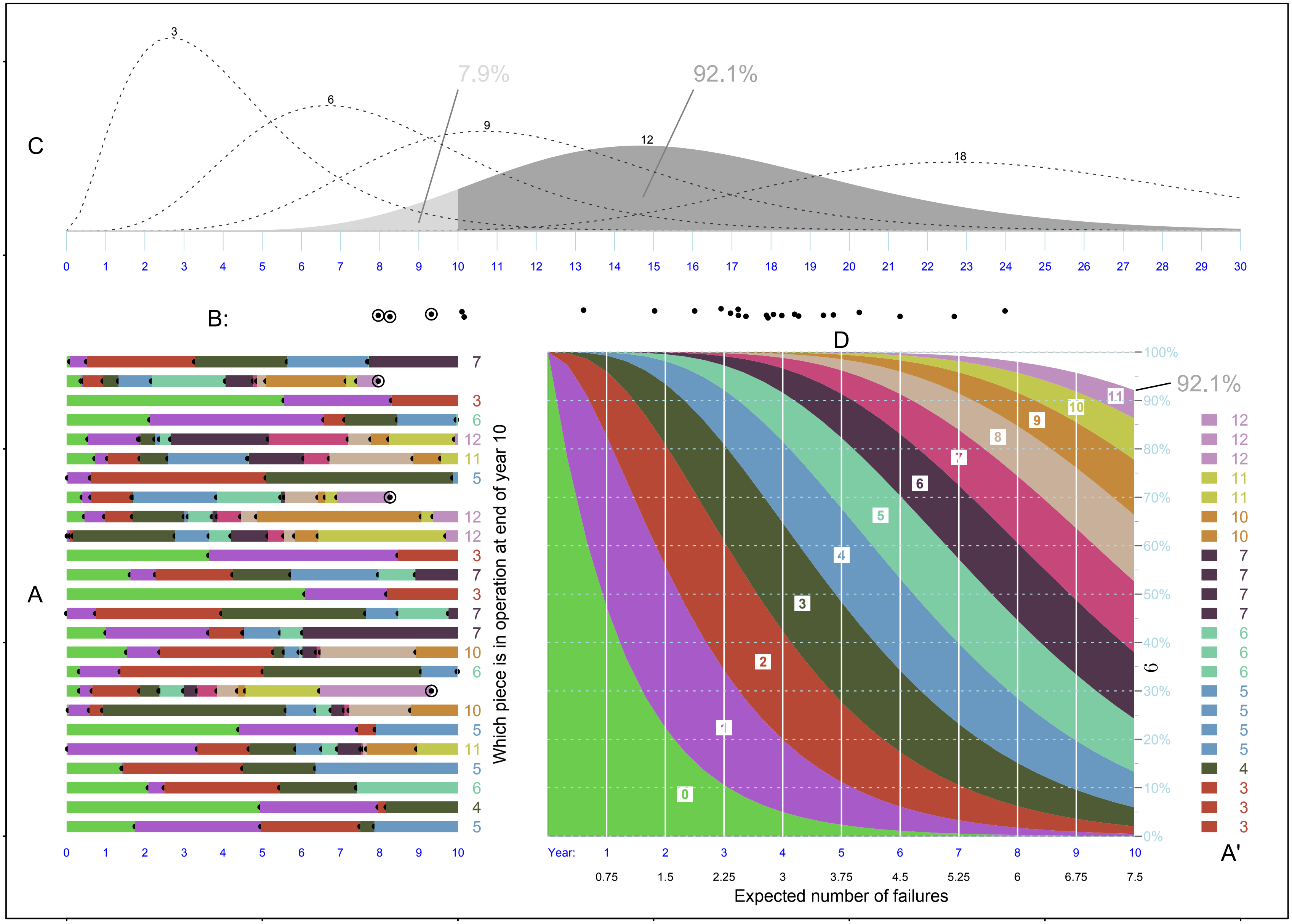
`http://www.r-project.org/`

It is compatible with UNIX operating systems, Windows and MacOS.

The `R` software is the collective work of the international statistical community. It is changing, thanks to the hundreds of statisticians and programmers who contribute constantly and on a voluntary basis to improve its content, including the addition of new modules that reflect the latest advances in statistical planning and data analysis. This software is widely used both for teaching and for research.

The code needed to produce the figure below is too complex to be reproduced here (we will provide it on request), but if you type ppois (11,7.5) or pgamma (10.12, scale = 4/3, lower. tail = FALSE) on the command line, the software gives 0.9207587 in both cases. Upon reading the article, can you say why?

For more information about the freeware `R` and its multiple uses, see for example the book entitled *Le logiciel `R` : Maîtriser le langage ? Effectuer des analyses statistiques*, published in 2011 by Springer France (ISBN 978-2-8178-0114-8). This beautiful book, a finalist for the Roberval Prize, is the work of three professors of statistics, Pierre Lafaye Micheaux (Université de Montréal), Rémy Drouilhet (Université Pierre Mendès France, Grenoble) and Benoît Liquet (The University of Queensland, Australia).


C
3
6
9
12
18
7.9%
92.1%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
B:
A
7
3
6
12
11
5
12
12
3
7
3
7
7
10
6
10
5
11
5
6
4
5
0 1 2 3 4 5 6 7 8 9 10
Which piece is in operation at end of year 10
D
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
92.1%
0
1
2
3
4
5
6
7
8
9
10
11
Year: 1 2 3 4 5 6 7 8 9 10
0.75 1.5 2.25 3 3.75 4.5 5.25 6 6.75 7.5
Expected number of failures
A'
12
12
12
11
11
10
10
7
7
7
7
6
6
6
5
5
5
5
4
3
3
3

from the most probable

$$P = \cdot5586.$$

In 56 cases out of a hundred such trials we should on a random selection get more improbable results than we have done. Thus we may consider the fit remarkably good.

Illustration V.

The following table gives the frequencies observed in a system recorded by Thiele in his *Forelaesinger over almindelig Iagttagelseslaere*, 1889, together with the results obtained by fitting a curve of my Type I. The rough values of the moments only were, however, used, and as well ordinates used measure areas :—

Groups.	Observed m'.	Curve m_1.	e.	e^2.	e^2/m.
1	0	·18	− ·18	·0324	·18
2	3	·68	− 2·32	5·3824	7·9153
3	7	13·48	+ 6·48	41·9904	3·1150
4	35	45·19	+10·19	103·8361	2·2977
5	101	79·36	−21·64	468·2896	5·9008
6	89	96·10	+ 7·10	50·4100	·5245
7	94	90·90	− 3·10	9·6100	·1058
8	70	71·41	+ 1·41	1·9881	·0278
9	46	48·25	+ 2·25	5·0625	·1049
10	30	28·53	− 1·47	2·1609	·0757
11	15	14·94	− ·06	·0036	·0002
12	4	6·96	+ 2·96	8·7616	1·2523
13	5	2·88	− 2·12	4·4944	1·5605
14	1	1·06	+ ·06	·0036	·0035
15	0	·34	+ ·34	·1156	·3400
16	0	·10	+ ·10	·0092	·0960
17	0	·00	+ 0	·0	·0
Total ...	500	500·36*	+ ·36	...	23·5000

Thus gives $\frac{1}{2}\chi^2 = 11\cdot75 = \eta$, say.

Then

$$P = e^{-\eta}\left(1 + \frac{\eta}{1} + \frac{\eta^2}{\underline{|2}} + \frac{\eta^3}{\underline{|3}} + \frac{\eta^4}{\underline{|4}} + \frac{\eta^5}{\underline{|5}} + \frac{\eta^6}{\underline{|6}} + \frac{\eta^7}{\underline{|7}}\right).$$

Substituting and working out we find

$$P = \cdot101 = \cdot1, \text{ say.}$$

Or, in one out of every ten trials we should expect to differ from the frequencies given by the curve by a set of deviations as improbable or more improbable. Considering that we should get a better fit of our observed and calculated frequencies by (i.) reducing the moments, and (ii.) actually

[*] Due to taking ordinates for areas and fewer figures than were really required in the calculations.

TABLE OF VALUES OF P FOR VALUES OF χ^2 and n' ; χ^2 from 1 to 70, n' from 3 to 20 *.

χ^2 \ n'	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
1	·606,531	·801,253	·909,796	·962,566	·985,612	·994,829	·998,249	·999,438	·999,828	·999,950	·999,986	·999,997	·999,999	1·	1·	1·	1·	1·
2	·367,879	·572,407	·735,759	·849,146	·919,699	·959,839	·981,012	·991,466	·996,340	·998,494	·999,406	·999,772	·999,917	·999.969	·999,990	·999,996	·999,999	·999,999
3	·223,130	·391,633	·557,825	·699,994	·808,847	·885,010	·934,357	·964,303	·981,424	·990,734	·995,544	·997,942	·999,074	·999,605	·999,830	·999,938	·999,972	·999,997
4	·135,335	·261,470	·406,006	·549,422	·676,676	·779,783	·857,123	·911,418	·947,347	·969,923	·983,436	·991,197	·995,466	·997,743	·998,903	·999,489	·999,763	·999,899
5	·082,085	·171,799	·287,298	·415,882	·543,813	·659,965	·757,576	·834,310	·891,178	·931,168	·957,979	·975,195	·985,813	·992,128	·995,754	·997,772	·998,860	·999,433
6	·049,787	·111,611	·199,148	·306,220	·423,190	·539,750	·647.232	·739,919	·815,263	·873,366	·916,082	·946,154	·966,491	·979,749	·988,095	·993,187	·996,197	·997,930
7	·030,197	·071,888	·135,888	·220,631	·320,847	·428,870	·536,632	·637,110	·725,544	·799,074	·857,613	·902,142	·934,711	·957,640	·973,260	·983,539	·990,125	·994.203
8	·018,316	·046,012	·091,578	·156,236	·238,103	·332,594	·433,470	·534,146	·628,837	·713,304	·785,131	·843,601	·889,327	·923,783	·948,867	·966,547	·978,637	·986,671
9	·011,109	·029,291	·061,099	·109,064	·173,578	·252,656	·342,296	·437,274	·532,104	·621,892	·702,931	·772,944	·831,051	·877,518	·913,414	·940,262	·959,743	·973,479
10	·006,738	·018,567	·040,428	·075,236	·124,652	·188,574	·265,026	·350,486	·440,493	·530,388	·615,960	·693,935	·762,183	·819,740	·866,628	·903,611	·931,906	·952,946
15	·000,553	·001,817	·004,701	·010,363	020,256	·036,000	·059,145	·090,810	·132,061	·182,371	·241,436	·307,227	·378,154	·450,691	·524,638	·594,156	·661,967	·721,272
20	·000,045	·000,170	·000,499	·001,250	·002,769	·005,570	·010,336	·017,913	·029,253	·045,341	·067,086	·095,212	·130,141	·171,934	·220,220	·274,231	·332,819	·394,580
25	·000,004	·000,016	·000,050	·000,139	000,341	·000,759	·001,554	·002,971	·005,345	·009,117	·014,822	·023,084	·034,566	·049.943	·069,824	·094,710	·124,915	·160,542
30	·000,000	·000,001	·000,005	·000,015	·000,039	·000,095	·000,211	·000,439	·000,857	·001,585	·002,792	·004,710	·007,632	·011,921	·018,002	·026,345	·037,446	·051,798
40	·000,000	·000,000	·000,000	·000,000	·000,000	·000,001	·000,003	·000,008	·000,017	·000,036	·000,072	·000,138	·000,255	·000,453	·000,778	·001,294	·002,087	·003,272
50	·000,000	·000,000	000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000,5	·000,001	·000,003	·000,006	·000,012	·000,023	·000,042	·000,075	·000,131
60	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,001	·000,002	·000,004
70	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000

[*] I have to thank Miss Alice Lee, D.Sc., for help in the calculation of part of this table. The certainty, *i. e.* the 1· in columns 16 to 20, denotes, of course, something greater than ·999,9995, *i. e.* unity to six figures.

The material on the previous page is taken from the 1900 article 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling' in Philosophical Magazine Series 5.

To Pearson, the $n' = 17$ groups implied a χ^2 distribution with what today we would call 16 'degrees of freedom.' So, he is interested in the quantity$P(\chi^2_{16} > 23.5)$. And, it seems he is not bothered by the fact that some of the expected (fitted) frequencies are quite small.

A normal approximation, with $\mu = 16$ and $\sigma = 32^{1/2}$ yields P = 0.092. A better one, due to Fisher, is that $(2\chi^2)^{1/2}$ is approximately normal with mean $(2df - 1)^{1/2}$ and unit variance.

Pearson's exact calculation of the upper tail, which is obvious now as equivalent to the lower tail of a Poisson distribution, yields P=0.101. The `pchisq(23.5,16,lower.tail=FALSE)` statement in `R` yields 0.1010081.

The tabulated P's have columns for n' values from 3 to 20. Here n' is the number of categories or 'groups.' Today we would call them 3 - 1 = 2 to 20 - 1 = 19 'degrees of freedom.'

Incidentally, in his paper he makes a distinction between fits to models with *known* parameter values (such as in gambling) and ones where, as in Illustration V, the parameters were *estimated* from the data. But he argued the correction for the altter would be small. The need to adjust the degrees of freedom to account for the number of fitting parameters became a *cause célèbre* between Pearson and Fisher (Fisher was correct).

As for the table of P values, the following R code gives the 'to 6 digits' P values we would obtain today.

```
n.prime = 3:20 ; length(n.prime)
chi.sq  = c(1:10,seq(15,30,5),seq(40,70,10))

P = matrix(NA,length(chi.sq),length(n.prime))

for(row in 1:length(chi.sq) ){
for(col in 1:length(n.prime) ){
       P[row,col] = pchisq(
             chi.sq[row], n.prime[col]-1,
             lower.tail = FALSE )
    }
}
cbind(chi.sq,round(P[,1:10],6))
cbind(chi.sq, round( P[,(14:18)] ,6) )
```

You can check whether Pearson and Lee made any errors.

In the Accromath article, we worked with the **gamma rather than the χ^2 distribution**, but Pearson's worked out P value shows the identity between the mass of the χ^2_{16} distribution to the right of the calculated statistic 23.5, and the sum of the first 8 terms (0 to 7) of the Poisson distribution with expectation $\mu = 23.5/2$. How is this?

Today, we know the link between the sum of the (0 to 7) probabilities of the Poisson distribution and the upper tail of the gamma(8) or Erlang(8) distribution. But at that time, even though the gamma function goes back to Euler, the gamma distributions had not been tabulated, and so for most of the 20th century statisticians relied instead on the exact link between the Poisson and the χ^2 distribution.

With tabulated distributions no longer critical, and where each distribution is covered in `R`, we tend not to rely on links. But, in any case let's complete these specific links.

A χ^2_{16} random variable is a sum of the squares of 16 independent Gaussian (Normal) random variables, each with mean 0 and standard deviation 1, i.e., $\chi^2_{16} = Z_1^2 + \ \ldots \ + Z_{16}^2$. The distribution of each Z^2 can easily be worked out from first principles. Ignoring the constant $(2\pi)^{1/2}$, the random variable Z has pdf

$$pdf_Z(z) \propto e^{-(1/2)z^2}$$

Ignoring the factor of 2 that reflects values coming from the squares of both negative and positive values of Z, we have

$$pdf_Y(y) \propto pdf_Z(\text{z-equivalent of } y) \times J(y),$$

where J is the Jacobian, $\frac{d\ y^{1/2}}{dy}$, evaluated at y. The pdf_Z(z-equivalent of y) is proportional to $e^{-(1/2)y}$ and $J(y) = (1/2)y^{-1/2}$ is the scaling factor.

So,

$$pdf_Y(y) \propto e^{-(1/2)y}\ y^{-1/2} = e^{-(1/2)y}\ y^{1/2-1}$$

and we recognize this as the pdf of a gamma distribution with *shape* parameter 1/2 and '*scale*' parameter 2, or '*rate*' parameter 1/2. [Check: we know that $\mathrm{E}(\chi^2_1) = \mathrm{E}(Z^2) = 1$, and that $\mathrm{Var}(\chi^2_1) = 2$. This fits with our knowledge that if G is a gamma random variable, then E(G) = $shape \ \times \ scale$ and Var(G) = $shape \ \times \ scale^2 = (1/2) \times 2^2 = 2$.]

We also know that the gamma family is closed under the addition of independent random variables with the same scale/rate, but possible different shapes. So the sum of two independent χ^2_1 distributions is a gamma random variable with shape parameter $2\times 1/2 = 1$ and 'scale' parameter 2, or 'rate' parameter 1/2. **In other words, the sum of the squares of 2 independent standardized Normal random variables has an exponential distribution with mean 2.**

So, the sum of the squares of 16 independent standardized Normal random variables has an gamma distribution with shape parameter 8 and scale 2. **Thus, *half* the sum of the squares of 16 independent standardized Normal random variables has an gamma distribution with shape parameter 8 and scale 1.** Thus, the fact that Pearson's calculation of the upper tail of the χ^2_{16} distribution looks identical to the calculation of the (0-7) lower tail area of a Poisson random variable is no longer a surprize.

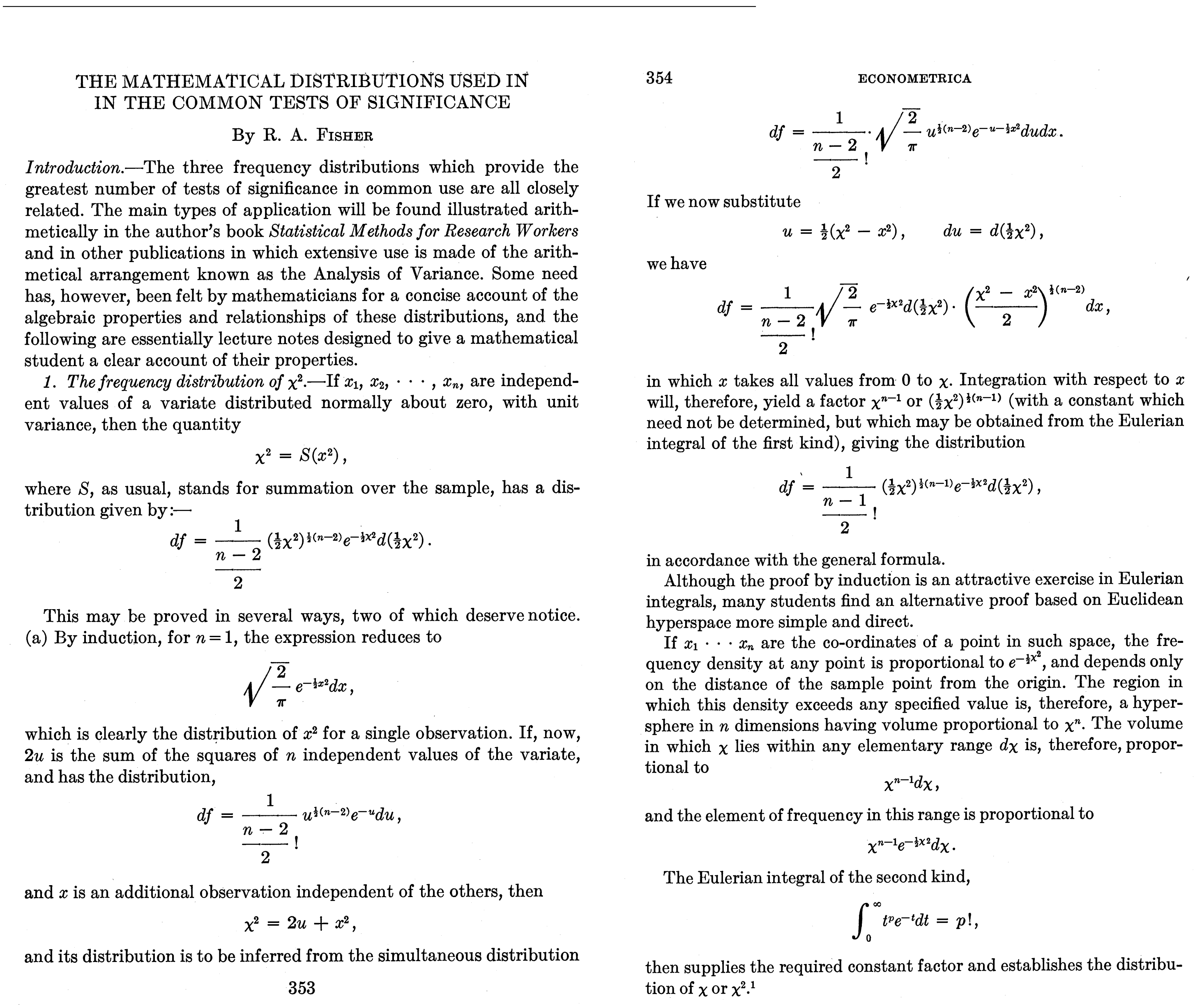
For most of the 20th century, before R, but after the χ^2 distributions had been tabulated, epidemiologists used the upper tail of the χ^2 to calculate the lower tail of the Poisson distribution, and to calculate exact confidence intervals for the expected value of a Poisson random variable. (See Fisher's example below) In effect, they were using the following links between these 3 random variables

- Poisson[$\mu = \lambda t$]
- Gamma[$shape,\ rate\ \lambda$]
- χ^2_{df}

$$\Pr(\text{ Poisson}[\mu]\ \leq\ k-1\ ) = \Pr(\text{ Gamma}[k,\lambda]\ > t\ ),$$

and

$$\Pr(\text{Gamma}[k,\lambda] > t) = \Pr\Big(\frac{\lambda}{2}\ \chi^2_{2k} > t\Big) = \Pr\Big(\chi^2_{2k} > \frac{2}{\lambda}\ t\Big).$$



THE MATHEMATICAL DISTRIBUTIONS USED IN IN THE COMMON TESTS OF SIGNIFICANCE

By R. A. Fisher

Introduction.—The three frequency distributions which provide the greatest number of tests of significance in common use are all closely related. The main types of application will be found illustrated arithmetically in the author's book *Statistical Methods for Research Workers* and in other publications in which extensive use is made of the arithmetical arrangement known as the Analysis of Variance. Some need has, however, been felt by mathematicians for a concise account of the algebraic properties and relationships of these distributions, and the following are essentially lecture notes designed to give a mathematical student a clear account of their properties.

1. The frequency distribution of χ^2.—If $x_1, x_2, \cdots, x_n$, are independent values of a variate distributed normally about zero, with unit variance, then the quantity

$$\chi^2 = S(x^2),$$

where S, as usual, stands for summation over the sample, has a distribution given by:—

$$df = \frac{1}{\frac{n-2}{2}}(\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n-2)}e^{-\frac{1}{2}\chi^2}d(\tfrac{1}{2}\chi^2).$$

This may be proved in several ways, two of which deserve notice. (a) By induction, for $n=1$, the expression reduces to

$$\sqrt{\frac{2}{\pi}}\,e^{-\frac{1}{2}x^2}dx,$$

which is clearly the distribution of x^2 for a single observation. If, now, $2u$ is the sum of the squares of n independent values of the variate, and has the distribution,

$$df = \frac{1}{\frac{n-2}{2}!}u^{\frac{1}{2}(n-2)}e^{-u}du,$$

and x is an additional observation independent of the others, then

$$\chi^2 = 2u + x^2,$$

and its distribution is to be inferred from the simultaneous distribution

$$df = \frac{1}{\frac{n-2}{2}!}\cdot\sqrt{\frac{2}{\pi}}\,u^{\frac{1}{2}(n-2)}e^{-u-\frac{1}{2}x^2}dudx.$$

If we now substitute

$$u = \tfrac{1}{2}(\chi^2 - x^2), \qquad du = d(\tfrac{1}{2}\chi^2),$$

we have

$$df = \frac{1}{\frac{n-2}{2}!}\sqrt{\frac{2}{\pi}}\,e^{-\frac{1}{2}\chi^2}d(\tfrac{1}{2}\chi^2)\cdot\left(\frac{\chi^2 - x^2}{2}\right)^{\frac{1}{2}(n-2)}dx,$$

in which x takes all values from 0 to χ. Integration with respect to x will, therefore, yield a factor χ^{n-1} or $(\frac{1}{2}\chi^2)^{\frac{1}{2}(n-1)}$ (with a constant which need not be determined, but which may be obtained from the Eulerian integral of the first kind), giving the distribution

$$df = \frac{1}{\frac{n-1}{2}!}(\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n-1)}e^{-\frac{1}{2}\chi^2}d(\tfrac{1}{2}\chi^2),$$

in accordance with the general formula.

Although the proof by induction is an attractive exercise in Eulerian integrals, many students find an alternative proof based on Euclidean hyperspace more simple and direct.

If $x_1 \cdots x_n$ are the co-ordinates of a point in such space, the frequency density at any point is proportional to $e^{-\frac{1}{2}\chi^2}$, and depends only on the distance of the sample point from the origin. The region in which this density exceeds any specified value is, therefore, a hypersphere in n dimensions having volume proportional to χ^n. The volume in which χ lies within any elementary range $d\chi$ is, therefore, proportional to

$$\chi^{n-1}d\chi,$$

and the element of frequency in this range is proportional to

$$\chi^{n-1}e^{-\frac{1}{2}\chi^2}d\chi.$$

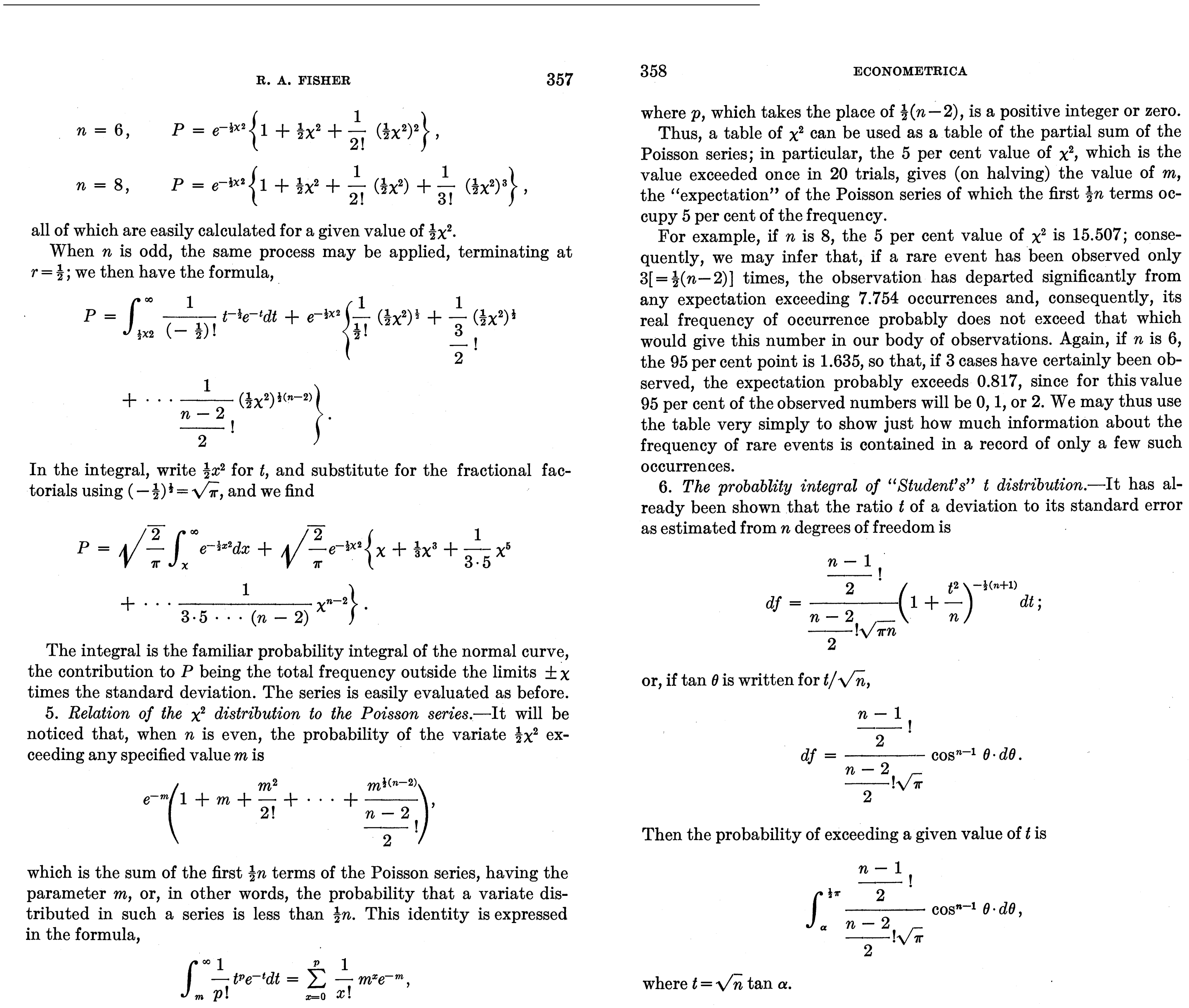
The Eulerian integral of the second kind,

$$\int_0^\infty t^p e^{-t}dt = p!,$$

then supplies the required constant factor and establishes the distribution of χ or χ^2.[1]

This article connected many distributions. And in section 5 (next page) it gave a way to use the χ^2 tables to obtain (without any trial and error) an exact confidence interval for the expectation of a Poisson random variable. Section 6 gives a similar CI, based on what we now call the F table, for the expectation of a Binomial random variable.

Fisher could think geometrically in n dimensions, and shorten some derivations to a few lines. just as he did when, still a student in 1912, he wrote to Pearson with one for of the distribution of the sample variance s^2 when $y_1, \ldots, y_n \sim N(\mu, 1)$, with μ unknown (thus $n-1$ *df*). His 'alternative proof' in this column (μ known, thus n *df*) uses the same insights.



R. A. FISHER 357

$$n = 6, \qquad P = e^{-\frac{1}{2}\chi^2}\left\{1 + \tfrac{1}{2}\chi^2 + \frac{1}{2!}\,(\tfrac{1}{2}\chi^2)^2\right\},$$

$$n = 8, \qquad P = e^{-\frac{1}{2}\chi^2}\left\{1 + \tfrac{1}{2}\chi^2 + \frac{1}{2!}\,(\tfrac{1}{2}\chi^2) + \frac{1}{3!}\,(\tfrac{1}{2}\chi^2)^3\right\},$$

all of which are easily calculated for a given value of $\frac{1}{2}\chi^2$.

When n is odd, the same process may be applied, terminating at $r=\frac{1}{2}$; we then have the formula,

$$P = \int_{\frac{1}{2}\chi 2}^{\infty} \frac{1}{(-\frac{1}{2})!}\,t^{-\frac{1}{2}}e^{-t}dt + e^{-\frac{1}{2}\chi 2}\left\{\frac{1}{\frac{1}{2}!}\,(\tfrac{1}{2}\chi^2)^{\frac{1}{2}} + \frac{1}{\frac{3}{2}!}\,(\tfrac{1}{2}\chi^2)^{\frac{1}{2}} + \cdots \frac{1}{\frac{n-2}{2}!}\,(\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n-2)}\right\}.$$

In the integral, write $\frac{1}{2}x^2$ for t, and substitute for the fractional factorials using $(-\frac{1}{2})^{\frac{1}{2}}=\sqrt{\pi}$, and we find

$$P = \sqrt{\frac{2}{\pi}}\int_{\chi}^{\infty} e^{-\frac{1}{2}x^2}dx + \sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\chi 2}\left\{\chi + \tfrac{1}{3}\chi^3 + \frac{1}{3\cdot 5}\,\chi^5 + \cdots \frac{1}{3\cdot 5 \cdots (n-2)}\,\chi^{n-2}\right\}.$$

The integral is the familiar probability integral of the normal curve, the contribution to P being the total frequency outside the limits $\pm\chi$ times the standard deviation. The series is easily evaluated as before.

5. *Relation of the χ^2 distribution to the Poisson series.*—It will be noticed that, when n is even, the probability of the variate $\frac{1}{2}\chi^2$ exceeding any specified value m is

$$e^{-m}\left(1 + m + \frac{m^2}{2!} + \cdots + \frac{m^{\frac{1}{2}(n-2)}}{\frac{n-2}{2}!}\right),$$

which is the sum of the first $\frac{1}{2}n$ terms of the Poisson series, having the parameter m, or, in other words, the probability that a variate distributed in such a series is less than $\frac{1}{2}n$. This identity is expressed in the formula,

$$\int_m^{\infty}\frac{1}{p!}\,t^pe^{-t}dt = \sum_{x=0}^{p}\frac{1}{x!}\,m^xe^{-m},$$

358 ECONOMETRICA

where p, which takes the place of $\frac{1}{2}(n-2)$, is a positive integer or zero.

Thus, a table of χ^2 can be used as a table of the partial sum of the Poisson series; in particular, the 5 per cent value of χ^2, which is the value exceeded once in 20 trials, gives (on halving) the value of m, the "expectation" of the Poisson series of which the first $\frac{1}{2}n$ terms occupy 5 per cent of the frequency.

For example, if n is 8, the 5 per cent value of χ^2 is 15.507; consequently, we may infer that, if a rare event has been observed only $3[=\frac{1}{2}(n-2)]$ times, the observation has departed significantly from any expectation exceeding 7.754 occurrences and, consequently, its real frequency of occurrence probably does not exceed that which would give this number in our body of observations. Again, if n is 6, the 95 per cent point is 1.635, so that, if 3 cases have certainly been observed, the expectation probably exceeds 0.817, since for this value 95 per cent of the observed numbers will be 0, 1, or 2. We may thus use the table very simply to show just how much information about the frequency of rare events is contained in a record of only a few such occurrences.

6. *The probablity integral of "Student's" t distribution.*—It has already been shown that the ratio t of a deviation to its standard error as estimated from n degrees of freedom is

$$df = \frac{\frac{n-1}{2}!}{\frac{n-2}{2}!\sqrt{\pi n}}\left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}(n+1)} dt;$$

or, if $\tan\theta$ is written for $t/\sqrt{n}$,

$$df = \frac{\frac{n-1}{2}!}{\frac{n-2}{2}!\sqrt{\pi}}\cos^{n-1}\theta\cdot d\theta.$$

Then the probability of exceeding a given value of t is

$$\int_{\alpha}^{\frac{1}{2}\pi}\frac{\frac{n-1}{2}!}{\frac{n-2}{2}!\sqrt{\pi}}\cos^{n-1}\theta\cdot d\theta,$$

where $t=\sqrt{n}\tan\alpha$.

The sums at the top of this column are the same ones Pearson had to wade through in 1900. The sum is messier when the df. is an odd number. Section 5 shows the link between the Poisson and χ^2 tail areas.

He obtains (without trial and error) an exact CI for the expectation of a Poisson random variable based on a count of 3. Today, trial and error is easy with the R ppois function. But the *principle* behind the exact CI remains unchanged.