**TRANSACTIONS OF SOCIETY OF ACTUARIES
1980 VOL. 32**

TESTING FOR SIGNIFICANT DIFFERENCES BETWEEN ACTUAL AND EXPECTED RESULTS

EDWARD J. SELIGMAN AND SHELDON KAHN*

ABSTRACT

The statistical technique of hypothesis testing is applied to the interpretation of actual-to-expected mortality and morbidity ratios. Exact and approximate methods are used for both homogeneous and heterogeneous populations. A format for a two-dimensional actual-to-expected report based on terminations from disability claim status is presented.

INTRODUCTION

THE comparison of actual results with expected results has long been a feature of retrospective statistical studies. In particular, the Society of Actuaries' mortality studies usually include a "mortality ratio," which is the ratio of actual deaths to expected deaths for a given age-sex class whose expected counts are based on a theoretical probability of death applied to the actual exposed lives. This concept has general application to other life or casualty contingencies if we think of both "actual" and "expected" as applying to the passage of an individual or entity from one state to another—for example, lives passing from life to death, from active status to disabled status, or from disabled status to active status or death, or dwellings passing from a state of being undamaged by fire to a state of being damaged by fire. The concept of an "actual-to-expected ratio" can be used in this general sense.

STATISTICAL BACKGROUND

Consider a population of n individuals in a given state, each with expected probability q_i $(i = 1, \ldots, n)$ of passing to another given state. The expected value of the number of individuals passing to the new state is

$$e = \sum_{i=1}^{n} q_i \,. \qquad (1)$$

* Mr. Kahn, not a member of the Society, is a senior analyst at CNA Insurance Company.

Suppose that we observe a individuals passing to the new state. The actual-to-expected ratio is a/e. The natural question is whether the ratio is significantly different from 1. In other words, is the actual value sufficiently different from the expected to make us doubt that our population of n individuals is behaving according to the given q_i's? One way to answer this question is to place it in the framework of the test of a statistical hypothesis. We state the "null hypothesis" (H_0) as follows:

H_0: *The given q_i's are the correct probabilities of termination for the n individuals.*

Assuming that H_0 is true, we then compute the probability P of observing at least a if $a \geq e$, or at most a if $a < e$ (we use the "one-tailed" test of H_0 because it seems the most appropriate one in this situation). By convention, if this probability is less than 0.05, we reject H_0 and say that the deviation between actual and expected is "significant" or "significant at the 5 percent level." If the probability is less than 0.01, we reject H_0 and say that the deviation is "highly significant" or "significant at the 1 percent level." For any other value of P, H_0 is not rejected. There is, however, no reason for the user to be bound by the 0.05 or the 0.01 level of significance if it is inappropriate in a given situation.

We compute P by classifying the q_i's into one of five cases.

CASE 1: $q_1 = q_2 = \ldots = q, n < 5$.

Under H_0, the actuals are distributed according to the binomial distribution, so that the desired probability is given by

$$P = \sum_{x=0}^{a} \binom{n}{x} q^x(1 - q)^{n-x} \quad \text{if } a < e$$

$$= \sum_{x=a}^{n} \binom{n}{x} q^x(1 - q)^{n-x} \quad \text{if } a \geq e\,.$$

CASE 2: $q_1 = q_2 = \ldots = q, n \geq 5$.

Again the actuals have a binomial distribution, but for large n the computation time may be too great. If $nq < 0.8$, use the Poisson Gram-Charlier approximation to the binomial distribution. If $nq \geq 0.8$, use the Camp-Paulson approximation to the binomial distribution. The maximum error of approximation is then less than 0.005. See the Appendix for a description of both of these approximations.

CASE 3: At least one q_i is distinct; n is small.[1]

[1] Whether n is "small" or "large" depends on the user, who must decide whether to forgo the exactness of case 3 methods for case 4 and case 5 methods, which are approximate but far easier to apply as n increases.

The actuals are distributed as the sum of distinct Bernoulli distributions (binomial with $n = 1$). We have a choice of using Waring's theorem [2, pp. 217–20] or one of several more recent methods [4], which require the use of a digital computer to obtain P exactly. However, the user should be aware of the numerical difficulties inherent in the use of either the latter methods or Waring's theorem (see Appendix).

CASE 4: At least one q_i is distinct, and the q_i's are neither all small (<0.05) nor all large (>0.95); n is large.

Use the central limit theorem. The normal distribution that approximates the sum of the Bernoulli distributions has mean $\Sigma_{i=1}^{n} q_i$ and variance $\Sigma_{i=1}^{n} q_i(1 - q_i)$.

CASE 5: At least one q_i is distinct, and all q_i's are small (<0.05). If all q_i's are large (>0.95), use symmetry to apply this case to the p_i's ($p_i = 1 - q_i$); n is large.

Use the Poisson approximation (see Appendix) to the sum of Bernoulli distributions. The mean is identical with the mean of the corresponding normal distribution for case 4.

Thus we can decide whether the actual outcome is significantly different from the expected outcome by testing H_0.

APPLICATION TO AN ACTUAL-TO-EXPECTED REPORT BASED ON DISABILITY INCOME TERMINATIONS

Consider a population of disabled claimants under disability income policies.

Each claimant is classified in two ways: by attained age and by duration of disablement. We can use a morbidity table such as the 1964 Commissioners Disability Table (1964 CDT) to compute the probability of termination within a given time period for any claimant. Table 1 shows probabilities of termination based on the 1964 CDT. For any mix of claimants classified by attained age and duration of disablement, we can calculate the expected number of terminations, e, from equation (1) and apply the test of H_0. The actual number of terminations, a, is then used along with e and the number of claimants, n, to compute P, the probability of observing the difference between a and e, the method used depending upon which of the cases 1–5 applies.

The quantity P is extremely important. It replaces subjective judgments about mortality ratios and morbidity ratios based on varying sample sizes by a single number that can be used to judge whether an actual result differs significantly from the expected result, or whether the difference is due to random fluctuation. In other words, this procedure "substitutes facts for appearances and demonstrations for impressions."

TABLE 1

PROBABILITIES OF TERMINATION BASED ON 1964 CDT

DURATION OF DISABILITY (MONTHS)	AGE AT DISABLEMENT								
	22	27	32	37	42	47	52	57	62
0......	0.750	0.751	0.737	0.717	0.687	0.653	0.615	0.569	0.522
1......	0.640	0.640	0.630	0.610	0.590	0.559	0.532	0.498	0.449
2......	0.530	0.529	0.519	0.500	0.480	0.450	0.414	0.369	0.315
3......	0.447	0.447	0.445	0.442	0.419	0.391	0.354	0.306	0.251
4......	0.371	0.369	0.368	0.366	0.348	0.322	0.289	0.245	0.199
5......	0.300	0.297	0.297	0.294	0.279	0.256	0.229	0.186	0.148
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
41.....	0.018	0.016	0.014	0.013	0.012	0.011	0.010	0.010	0.010
42.....	0.017	0.016	0.014	0.013	0.011	0.010	0.010	0.009	0.009
43.....	0.017	0.015	0.014	0.012	0.011	0.010	0.010	0.009	0.009

TWO-DIMENSIONAL ACTUAL-TO-EXPECTED REPORT

Table 2 is a reproduction of a two-dimensional actual-to-expected termination report. The first dimension of the report is real time (calendar period), which increases in the horizontal direction from right to left across the page. The second dimension of the report is duration of disablement upon termination, which increases in the vertical direction down the page. Note that both dimensions are expressed in terms of time, in contrast to the conventional multidimensional report, where one dimension might be calendar time and another might be type of disablement. As we point out later, the use of time for both dimensions gives the report a certain advantage.

The table indicates, by means of an asterisk (*) or a double asterisk (**), those cells for which the actual terminations are significantly different from the expected terminations at the 5 percent and 1 percent levels, respectively. An actual-to-expected report in this format enables the user to spot significant trends in termination from disability without being diverted by deviations that can be attributed to random fluctuations.

The user should, however, be aware that, in an actual-to-expected report with k cells, one can expect $(0.05 - 0.01)k = 0.04k$ cells with a single asterisk, and $0.01k$ cells with a double asterisk, even if the claim population does terminate in accordance with the given morbidity table. This is because the hypothesis-testing procedure implies that at the 5 percent level of significance we reject H_0 5 percent of the time even when H_0 is,

TABLE 2

Actual Terminations (A) and Actual-to-Expected Ratios (A/E)

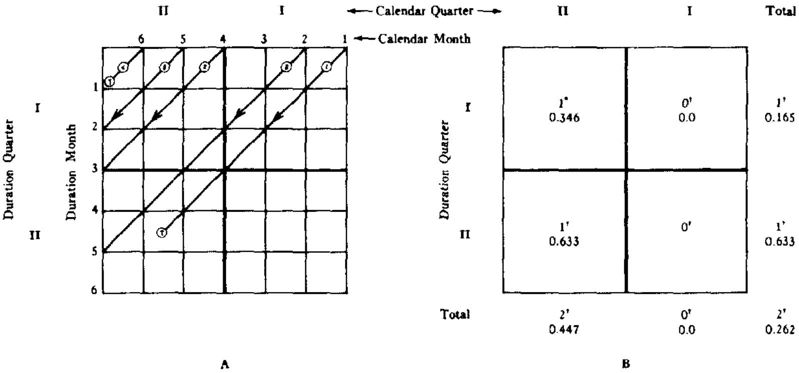
Duration of Disablement before Termination (Months)	Calendar Period				Total
	1978	1977	1976	1975	
1–2:					
A	57	84	211	375	727*
A/E	0.96	0.97	0.95	0.95	0.95
2–3:					
A	86**	89**	202**	353	730**
A/E	0.83	0.71	0.85	0.98	0.88
3–4:					
A	106	120**	208	317	751**
A/E	0.89	0.84	0.96	0.99	0.94
4–5:					
A	87**	99**	147**	211**	544**
A/E	1.28	1.21	1.17	1.19	1.20
5–6:					
A	58**	77**	101	155**	391**
A/E	1.26	1.36	1.06	1.23	1.21
6–9:					
A	174	238	318	385**	1,115**
A/E	0.98	1.08	1.04	1.19	1.09
9–12:					
A	99	122	159	184**	564
A/E	1.06	0.98	0.98	1.18	1.05
12–15:					
A	50	92	101	101	344
A/E	0.85	1.04	0.97	1.05	0.99
15–18:					
A	38	58	59*	53*	208**
A/E	0.89	0.87	0.83	0.83	0.85
18–24:					
A	51*	76**	77	64**	268**
A/E	0.78	0.80	0.86	0.79	0.81
24–36:					
A	72**	102	93	69	336
A/E	0.76	1.00	1.03	0.95	0.93
36–48:					
A	52	45	32	25*	154**
A/E	0.95	0.90	0.80	0.75	0.86
48–60:					
A	35	35*	23	19	112*
A/E	1.08	1.32	1.06	1.07	1.14
60–84:					
A	31	35*	22	19	107
A/E	0.99	1.34	0.94	0.99	1.07
84–120:					
A	25	19	18	17	79
A/E	1.06	0.95	0.92	0.76	0.92
Total:					
A	1,021*	1,291	1,771	2,347**	6,430
A/E	0.95	0.98	0.97	1.04	0.99

* Indicates that the difference between actual and expected is significant at the 5% level.

** Indicates that the difference between actual and expected is significant at the 1% level.

in fact, true. Conversely, we may fail to reject a false H_0 simply because there is not enough information available to justify the rejection. The Appendix gives a test of the entire actual-to-expected report that will tell the user whether he is justified in examining individual cells of the report for significant deviation between actual and expected.

Figure 1, *A*, is based on the accompanying list of hypothetical claims (see Table 3) and shows how claim records are used to build the two-dimensional actual-to-expected report. A newly disabled claimant starts at the top of the report and proceeds diagonally downward and leftward, contributing expected terminations to the appropriate cells, row sums, and column sums, until he terminates or passes off the leftmost edge of the report.



	II	I	Total
I	1* 0.346	0† 0.0	1† 0.165
II	1† 0.633	0†	1† 0.633
Total	2† 0.447	0† 0.0	2† 0.262

FIG. 1.—See text for explanation. Circled numbers identify claimants; a circled *T* at the end of diagonal indicates termination. An asterisk (*) indicates significance at the 5 percent level; a dagger (†) indicates significance level not calculated.

The individual cells in Figure 1, *A*, represent months on both the calendar-period and duration axes. However, the actual-to-expected report (Fig. 1, *B*) is based on quarterly periods in both directions, in order to illustrate better the hypothesis tests that are applied. The individual diagonal lines represent the passage of the individual claimant. Claimants who terminate are denoted by a *T* at the end of their diagonals.

The two-dimensional actual-to-expected report is a valuable tool for the analysis of disability income claim termination experience because the user can

1. Scan across the report to detect trends in morbidity ratios by calendar period.
2. Scan down the report to detect trends in morbidity ratios by duration of claim before termination.

3. Scan diagonally downward and leftward to follow, in a general way, the experience of a cohort of claimants. This is a direct result of the use of time for each of the two dimensions of the report.

All three of the above capabilities, in combination with the ability to identify and ignore nonsignificant morbidity ratios, have proved most useful in the analysis of actual-to-expected reports.

TABLE 3

APPLICATION OF HYPOTHESIS TESTING TO ACTUAL-TO-EXPECTED RATIOS: EXAMPLE

HYPOTHETICAL CLAIM RECORDS FOR FIGURE 1

Claimant	Age at Time of Claim	Calendar Month of Claim	Calendar Month Terminated
1.	32	1	5
2.	42	4	. . . .
3.	42	2	. . . .
4.	47	6	6
5.	52	5	. . . .

CONTRIBUTIONS TO EXPECTED TERMINATIONS

(Based on 1964 CDT)

Cell	Claimant	Contribution to Expected Terminations
Duration Quarter I, Calendar Quarter II.	2	0.687, 0.590, 0.480
	3	0.480
	4	0.653
	5	0.615, 0.532
Calendar Quarter I, all duration quarters.	1	0.737, 0.630, 0.519
	3	0.687, 0.590
Duration Quarter II, all calendar quarters.	1	0.445, 0.368
	3	0.419, 0.348

TEST OF SIGNIFICANCE OF DIFFERENCE BETWEEN ACTUAL AND EXPECTED TERMINATIONS FOR CALENDAR QUARTER II, DURATION QUARTER I, USING WARING'S THEOREM

$q_1 = 0.687$, $q_2 = 0.590$, $q_3 = 0.480$, $q_4 = 0.480$, $q_5 = 0.653$, $q_6 = 0.615$, $q_7 = 0.532$

Actual terminations $= a = 1$

Expected terminations $= e = 0.687 + 0.590 + \ldots + 0.532 = 4.037$

Prob (0 or 1 termination) $= 1 -$ Prob (2 or more terminations)

$= 1 - 0.976 = 0.024$

Significant at the 5% level but not at the 1% level

MEANING OF "EXPECTED TERMINATIONS"

The phrase "expected terminations" is ambiguous. It is important that the user recognize that there are at least two reasonable interpretations of the term.

Interpretation 1 (prospective).—This is the viewpoint of a user who, on New Year's morning of, say, 1980, estimates the total terminations from disability claim status that will occur during 1980, on the basis of the beginning 1980 claim population and forecasts of claim incidence during 1980. The estimate is made at this time only and is not updated during the year. The expected terminations during 1980 are computed in two steps. First, all claimants as of midnight on December 31, 1979, are classified according to attained age and duration of disability. For each individual, the probability of terminating during the next year is calculated from the appropriate morbidity table. The sum of these probabilities is the first component of the expected terminations for 1980. Second, an incidence table is applied to the active insured lives and expected insured lives to estimate the newly disabled claimants who will appear during 1980. These lives also will be classified according to attained age, and assigned a duration of zero. The sum of their probabilities of termination during 1980 will be added to the first component of expected terminations during 1980. This sum is the "expected terminations" during calendar year 1980. It is a forecast of future results.

Interpretation 2 (retrospective).—This is the viewpoint of a user who, just before midnight on December 31, 1980, uses each of the beginning monthly claim populations of 1980 to estimate the total terminations from disability claim status that should have occurred during that month. The 1980 "expected terminations" is the sum of these monthly expected terminations for 1980. At the end of calendar year 1980, we examine the claim populations at the beginning of each calendar month during the year, and classify each according to attained age and duration of disability. The probability of terminating during that month is calculated from the appropriate morbidity table. The sum of these probabilities over all months of 1980 is the "expected terminations" during calendar year 1980. It is not a forecast of future results but rather a statement of what should have occurred during 1980.

Interpretation 2 (retrospective) generates month by month a set of conditional probabilities of termination. Since the sum of the probabilities is likely to exceed 1, a single claimant can generate expected terminations exceeding 1 over a quarter or year. This apparent anomaly is outweighed by the fact that a claim population that obeys the termination rates in the experience morbidity table will generate an actual-to-expected report with all the actual-to-expected ratios equal to 1.00. Because of this desirable feature, interpretation 2 was used for the actual-to-expected reports discussed in this paper.

ACKNOWLEDGMENTS

The authors wish to thank James R. Porter, A.S.A., for his expert programming assistance, and Philip E. Heckman, A.C.A.S., for several valuable discussions.

APPENDIX

All numerical examples are based on probabilities of termination found in the 1964 Commissioners Disability Table and shown in Table 1 of the paper.

CASE 1

Use the binomial distribution.

Example: Four claimants are all aged 47, disabled for two months. During the third month of disablement, two claimants terminate.

$$q_1 = q_2 = q_3 = q_4 = 0.450\,, \qquad q = 0.450\,;$$

$$P = \sum_{x=2}^{4} \binom{4}{x} 0.45^x 0.55^{4-x} = 0.609 \text{ (not significant)}\,;$$

H_0 is not rejected .

CASE 2

The references here are Raff [3] and Gebhardt [1]. The maximum error of approximation can be kept to 0.005 if we follow the rule already given, that is, use Camp-Paulson if $nq \geq 0.8$, otherwise use Poisson Gram-Charlier. Note that the first formula on page 1644 of Gebhardt's paper is in error. It should read

$$B_7(k, n, p) = R(k, np) + 0.5p(k - np)r(k, np)\,.$$

The Camp-Paulson approximation to the cumulative binomial distribution, that is, to

$$\sum_{x=0}^{a} \binom{n}{x} q^x (1 - q)^{n-x}\,,$$

is given by the following:

Let

$$y = \left[\frac{(n - a)q}{(a + 1)(1 - q)}\right]^{1/3} \left(9 - \frac{1}{n - a}\right) + \frac{1}{a + 1} - 9\,.$$

Let

$$z = \frac{1}{n - a} \left[\frac{(n - a)q}{(a + 1)(1 - q)}\right]^{2/3} + \frac{1}{a + 1}\,.$$

Then

$$P = \phi\left(\frac{-y}{3z^{1/2}}\right) \approx \sum_{x=0}^{a} \binom{n}{x} q^x (1 - q)^{n-x}\,,$$

where

$$\phi(v) = (2\pi)^{-1/2} \int_{-\infty}^{v} \exp(-t^2/2)dt$$

(the cumulative distribution function of the standardized normal distribution). The Poisson Gram-Charlier approximation to the cumulative binomial distribution is given by

$$P = \sum_{x=0}^{a} \frac{\exp(-nq)(nq)^x}{x!} + 0.5q(a - nq) \frac{\exp(-nq)(nq)^a}{a!} .$$

Example: Forty claimants are all aged 22, disabled for five months. During the sixth month of disablement, seven claimants terminate. We have $q_1 = q_2 = \ldots = q_{39} = q_{40} = 0.300$, and $q = 0.300$, $n = 40$, $a = 7$. Since $nq \geq 0.8$, we use the Camp-Paulson approximation, which gives $P = 0.0552$ (not quite significant at the 5 percent level). The Poisson Gram-Charlier approximation gives $P = 0.0567$. The normal curve approximation using the central limit theorem yields $P = 0.0423$. The answer, exact to 4 decimal places, is $P = 0.0553$. While the Camp-Paulson approximation is appropriate here, the Poisson Gram-Charlier approximation is quite good, considering the large value of q.

CASE 3

References [2] and [4] discuss methods appropriate for an exact computation when the termination rates are heterogeneous. These methods, however, all require the summation of many terms of varying magnitude. It is imperative that double-precision arithmetic be used when these summations are performed, in order to avoid answers that are completely erroneous. The fact that floating-point addition is not associative is too often ignored. To convince oneself of this fact, consider the quantities $x_1 = 0.1$, $x_2 = x_3 = \ldots = x_{100} = 0.00001$. The two sums $x_1 + (x_2 + x_3 + \ldots + x_{100})$ and $x_1 + x_2 + \ldots + x_{100}$ are algebraically identical, but using single precision on a computer with a 32-bit word length (24 bits for floating-point mantissa) yields 0.1009900 for the first sum and 0.1009855 for the second. If the second sum is done using double-precision arithmetic, the result is 0.1009900, which is correct to 7 significant figures. Double-precision arithmetic is almost always enough to give good answers for summations of many quantities. We could also separate plus and minus terms, sort in order of magnitude, and add them in increasing numerical order, but this requires additional machine time. Since most present-day computers use little additional machine time and almost no extra programming time for double-precision arithmetic, the authors urge that double-precision arithmetic be the standard for all floating-point computations.

Example: Four claimants are aged 27, 32, 47, and 62, disabled for three, one, four, and two months, respectively. During the next month of disablement, three claimants terminate. Use Waring's theorem [2]:

$$q_1 = 0.447\,, \quad q_2 = 0.630\,, \quad q_3 = 0.322\,, \quad q_4 = 0.315\,;$$

$$\begin{aligned} P &= \text{Prob (3 or more terminations)} \\ &= 0.447 * 0.630 * 0.322 + 0.447 * 0.630 * 0.315 \\ &\quad + 0.447 * 0.322 * 0.315 + 0.630 * 0.322 * 0.315 \\ &\quad - 3 * 0.447 * 0.630 * 0.322 * 0.315 \\ &= 0.203 \text{ (not significant)}\,. \end{aligned}$$

Case 4

Use the central limit theorem.

Example: There are twenty-seven claimants, three from each of the nine age classes, disabled for two months. During the next month of disablement, eighteen claimants terminate. The mean of the normal distribution is given by $3(0.530 + 0.529 + 0.519 + \ldots + 0.315) = 12.318$, while the variance is $3(0.530 * 0.470 + 0.529 * 0.471 + \ldots + 0.315 * 0.685) = 6.558$, and the standard deviation is 2.561. Then,

$$\text{Prob } (a > 17.5) = \text{Prob } [(a - 12.318)/2.561 > 2.023] = 1 - \phi(2.023)\,,$$

where

$$\phi(v) = (2\pi)^{-1/2} \int_{-\infty}^{v} \exp(-t^2/2)dt$$

(the cumulative distribution function of the standardized normal distribution). Since $(a - 12.318)/2.561$ is distributed as a normal random variable with mean 0 and variance 1, the observed value of a is 2.023 standard deviations away from its mean. The probability can be found from a table of normally distributed random variables and is 0.022. Thus, the result is significant at the 5 percent level but not at the 1 percent level.

Unfortunately, there seems to have been no work done in estimating an upper bound to the error of approximation made by using the central limit theorem in this case of heterogeneous probabilities.

Case 5

We justify the use of the Poisson distribution as an approximation to the sum of Bernoulli distributions by the following line of reasoning:

The moment generating function of a Bernoulli random variable is

$$M_X(t) = (1 - q) + q \exp(t) \, .$$

Thus the moment generating function of the sum of n such random variables, each with parameter q_i, is

$$M_S(t) = \prod_{i=1}^{n} \{1 + q_i [\exp(t) - 1]\} \, , \tag{2}$$

where $S = \Sigma_{i=1}^{n} X_i$. Rewrite equation (2) as

$$M_S(t) = \exp \left[\!\left[\sum_{i=1}^{n} \ln \{1 + q_i [\exp(t) - 1]\} \right]\!\right] ,$$

and, since the q_i's are small, replace the logarithm term by the first nonzero term of its Maclaurin series, resulting in

$$M_\lambda(t) \approx \exp \{[\exp(t) - 1]\lambda\} \, .$$

This is exactly the moment generating function of a Poisson random variable with mean and variance $\lambda = \Sigma_{i=1}^{n} q_i$.

Example: There are nine claimants, one from each of the nine age classes, each disabled for forty-three months. During the forty-fourth month of disablement, two claimants terminate. The Poisson random variable has parameter (mean and variance) equal to $0.017 + 0.015 + 0.014 + \ldots + 0.009 = 0.107$. The probability that $a \geq 2$ is

$$P = 1 - \sum_{x=0}^{1} [\exp(-0.107) 0.107^x]/x! = 0.00523 \, ,$$

which is significant at the 1 percent level.

Just as in case 4, there appears to be no literature on an estimate for the upper bound to the error of approximation in this case of heterogeneous probabilities.

TESTING THE ENTIRE ACTUAL-TO-EXPECTED REPORT

Suppose the entire actual-to-expected report has k cells, and actual terminations for each cell occur according to expected terminations. We would expect to find $(0.05 - 0.01)k$ cells with one asterisk, and $0.01k$ cells with two asterisks. We can use the chi-square test to tell us whether the actual number of no-asterisk, one-asterisk, and two-asterisk cells is different enough from the expected number of such cells to cause us to reject the null hypothesis (H_0') given by

H_0': *The expected terminations for all cells in the actual-to-expected report are the correct ones for the population under study.*

The table is shown below:

		*	**
Actual	a_0	a_1	a_2
Expected	e_0	e_1	e_2

We compute the usual Pearson chi-square statistic as follows:

$$\chi^2 \approx \sum_{i=0}^{2} \frac{(a_i - e_i)^2}{e_i}$$

with 2 degrees of freedom.

For example, Table 2 has 38 cells with no asterisks, 6 cells with one asterisk, and 16 cells with two asterisks. The table is

		*	**
Actual	38	6	16
Expected	57	2	1

The chi-square statistic is

$$\chi^2 \approx \frac{(38 - 57)^2}{57} + \frac{(6 - 2)^2}{2} + \frac{(16 - 1)^2}{1} = 239.3\,,$$

which is significant at the 1 percent level. Only now are we safe in examining individual cells of the table and making conclusions about whether actual and expected counts differ significantly with respect to the absence or presence of asterisks.

REFERENCES

1. Gebhardt, F. "Some Numerical Comparisons of Several Approximations to the Binomial Distribution," *Journal of the American Statistical Association*, LXIV (1969), 1638–46.
2. Jordan, C. W., Jr. *Life Contingencies*. Chicago: Society of Actuaries, 1967.
3. Raff, M. S. "On Approximating the Point Binomial," *Journal of the American Statistical Association*, LI (1956), 293–303.
4. White, R. P., and Greville, T. N. E. "On Computing the Probability that Exactly *k* of *n* Independent Events Will Occur," *TSA*, XI (1959), 88–99.

DISCUSSION OF PRECEDING PAPER

WILLIAM H. WOODALL:

The probability P is correctly used by the authors to determine whether the null hypothesis, H_0, should be accepted or rejected. However, the suggested values of 0.05 and 0.01 are not the levels of significance for the tests. This results from the fact that the proposed tests are not one-tailed tests. They are, in fact, two-tailed tests.

The level of significance of a test is the probability of rejecting H_0 when H_0 is true. If we define a' to be the random variable that takes the observed value a, then by their rules H_0 is rejected if and only if

$$P(a' > a) \leq \alpha$$

or

$$P(a' < a) \leq \alpha \, ,$$

where these probabilities are calculated under H_0 and $\alpha = 0.05$ or 0.01. But this rule implies that H_0 is rejected with probability 2α when H_0 is true. It is incorrect statistical reasoning to let the location of the rejection region of a test depend upon the observed value of the test statistic. A one-tailed test is appropriate in this case only when the direction of the difference between the actual and expected results is specified a priori.

For a level of significance of α, the P-values for the tests should be compared not to α but to $\alpha/2$. This rule change reverses the conclusion drawn in case 5.

(AUTHORS' REVIEW OF DISCUSSION)

EDWARD J. SELIGMAN AND SHELDON KAHN:

Mr. Woodall's point is valid in the context of what is known as a "prospective" experiment. For example, an experimenter who wishes to test the effect of a chemical fertilizer on crop yield would construct his test before observing the experimental outcome. On the other hand, consider the following "retrospective" experiment. A gambler is engaged in a coin-tossing game where he wins a dollar for every head that appears and his opponent wins a dollar for every tail. After 100 tosses, he finds that 80 tosses have landed tails. At this point, the gambler certainly will ask a question based on the null hypothesis that the coin is unbiased. One question might be, "What is the probability of observing exactly 80 tails in 100 tosses?" Another is, "What is the probability of observing

80 or more tails in 100 tosses?" We believe that the gambler has a right to be impatient with the statistician who refuses to answer any questions at all because they were not asked before the experiment began. It is the statistician's duty to explain that the first question may be relevant for a prospective experiment but is irrelevant for this retrospective experiment. We believe, however, that the second question is relevant for both prospective and retrospective experiments of the coin-tossing variety.

This is exactly what we are doing in each of cases 1–4, where we observe an experimental outcome based on disability income terminations. The actuary encounters principally the outcomes of retrospective experiments simply because he is seldom able to control the experimental conditions. Further, if the experimental outcome is affected by economic conditions, these conditions may not become known until the experiment is well under way or even completed.

However, for the actual-to-expected report (Table 2), we used a two-tailed test of the null hypothesis for each cell rather than the one-tailed test used in cases 1–4. The critical region for each test was set before the outcome was observed. Our reasoning was that we have a set of outcomes rather than a single result selected in advance. The question asked here was, "Does this cell represent a significant deviation from the expected result?" It is obviously very important that the experimenter realize what the null hypothesis and question are, and then construct the statistical test to be consistent with the question.