The Statistician (1995)
44, *No.* 4, *pp.* 417–423

Mid-*P* confidence intervals: a brief review

By G. BERRY†

University of Sydney, Australia

and P. ARMITAGE

University of Oxford, UK

[Received March 1995. Revised June 1995]

SUMMARY
Significance tests that are based on discrete probabilities are conservative in that the average value of the significance level, when the null hypothesis is true, always exceeds 0.5. An approach suggested by H. O. Lancaster over 40 years ago overcomes this problem. This is to calculate the mid-*P* value, where only half of the probability of the observed sample is included in the tail. The average value of the mid-*P* value is 0.5 and the variance is slightly less than that of a random variable uniformly distributed between 0 and 1. The mid-*P* concept has usually been advocated in the context of significance testing but it can be extended to the calculation of confidence intervals in an estimation approach by defining, for example, the 95% mid-*P* confidence limits as the values that have a one-sided mid-*P* value of 0.025. In this paper we review recent work supporting this approach.

Keywords: Confidence intervals; Discrete data; Exact tests; Mid-*P*

1. Introduction

When inferences are drawn from data arising as counts a problem, which has been long recognized, is that a significance test does not give a significant result at a specified level with the probability equal to that level when the null hypothesis is true; for example, '$P < 0.05$' will not be found for exactly 5% of random samples drawn from a population in which the null hypothesis is true.

In general, consider a discrete variable r, taking only integer values, and suppose that within the context of the situation r must be within the range $l \leqslant r \leqslant u$. This general situation includes the binomial distribution with $l = 0$ and $u = n$, a Poisson count with $l = 0$ and $u = \infty$, and a cell in a 2×2 table with l and u dependent on the marginal totals of the table. When the null hypothesis, that a parameter θ defining the distribution takes the value θ_0, is true then denote the probability of obtaining the observation by p_r, and define the one-sided significance level P as

$$P(r, \theta_0) = \sum_{i=l}^{r} p_i. \tag{1}$$

As an example, consider an observation r from a binomial distribution with sample size $n = 10$ and probability θ, and suppose that a test is required of the null hypothesis that $\theta = 0.5$, against the alternative hypothesis that $\theta < 0.5$ or $\theta > 0.5$. Then significance at the one-sided 2.5% level is found, for low values of r, only for $r = 0$ or $r = 1$, and for a two-sided alternative at the 5% level for $r = 0, 1, 9$ or 10. The probability of one or other of these values is 0.022. Therefore, a result that is significant at the 5% level would be found in only 2.2% of random samples if the null hypothesis

†*Address for correspondence*: Department of Public Health and Community Medicine, Edward Ford Building (A27), University of Sydney, Sydney, NSW 2006, Australia.
E-mail: geoffb@pub.health.su.oz.au

 0039–0526/95/44417

were true. In one sense this causes no difficulty if the precise level of P is stated. Thus if $r = 1$ we have that $P = 0.022$, and a result that is significant at a level of 0.022 or less would occur in exactly 2.2% of random samples. As it is impossible to construct a satisfactory test with the feature that a significant result at the 5% level, or any other prespecified level, will in general occur with a probability of 5% when the null hypothesis is true, quoting the precise level is the most satisfactory course, and there is no need to specify in advance any particular value as a boundary between significance and non-significance.

2. Mid-*P* tests

There remains, however, another feature which suggests that the usual form of test is unduly conservative. The expectation of P is given by

$$\begin{aligned} E(P) &= \sum_{r=l}^{u} \left(p_r \sum_{i=l}^{r} p_i \right) \\ &= \tfrac{1}{2} \left\{ \left(\sum_{r=l}^{u} p_r \right)^2 + \sum_{r=l}^{u} p_r^2 \right\} \\ &= \tfrac{1}{2} + \tfrac{1}{2} \sum_{r=l}^{u} p_r^2. \end{aligned} \tag{2}$$

This clearly exceeds $\frac{1}{2}$. For a significance test of a continuous variable, such as a test based on the normal distribution or the t-distribution, the significance level would be uniformly distributed over the range from 0 to 1 and so have a mean of $\frac{1}{2}$ and a variance of 1/12, and it is desirable that the significance level from a discrete variable should share these properties as closely as possible.

Lancaster (1949) considered the problem from the point of view of combining probabilities from different experiments by using the transformation $\chi^2 = -2 \ln P$ as advocated by Fisher (1932). Lancaster noted that there was a bias using the usual definition of P-values and suggested using the median value of the χ^2 defined as the transform of $P_m(r)$ where

$$P_m(r) = \tfrac{1}{2} \{ P(r, \theta_0) + P'(r, \theta_0) \}, \tag{3}$$

and P' is the probability of the more extreme observations only. Lancaster (1952) considered this further for Poisson counts referring to the probability as the *median probability*. Nine years later Lancaster (1961) repeated this definition referring to it as 'the median probability, or perhaps more appropriately the mid-probability'. The latter, abbreviated to mid-P, has become the usual terminology. The uniformly most powerful unbiased test is obtained through use of an auxiliary random experiment (Tocher, 1950). Lancaster (1961) noted that this is unsatisfactory in practice since it does not always give the same result and leads to inconsistencies but noted that the median probability agrees most often with the auxiliary random sampling method (see also Haber (1986)).

An alternative expression for the mid-P significance level is

$$P_m(r) = \sum_{i=l}^{r} p_i - \tfrac{1}{2} p_r . \tag{4}$$

The expectation of the mid-P level is

$$\begin{aligned} E(P_m) &= \sum_{r=l}^{u} \left\{ p_r \left(\sum_{i=l}^{r} p_i - \tfrac{1}{2} p_r \right) \right\} \\ &= \tfrac{1}{2} \left(\sum_{r=l}^{u} p_r \right)^2 \\ &= \tfrac{1}{2}. \end{aligned} \tag{5}$$

The variance of the mid-P level is

$$\begin{aligned}\text{var}(P_m) &= \sum_{r=l}^{u}\left\{p_r\left(\sum_{i=l}^{r} p_i - \tfrac{1}{2}p_r\right)^2\right\} - E(P_m)^2 \\ &= \tfrac{1}{3}\left(\sum_{r=l}^{u} p_r\right)^3 - \tfrac{1}{12}\sum_{r=l}^{u} p_r^3 - \tfrac{1}{4} \\ &= \tfrac{1}{12}\left(1 - \sum_{r=l}^{u} p_r^3\right). \end{aligned} \tag{6}$$

Thus the mid-P value has the same mean and a slightly smaller variance than a variable uniformly distributed between 0 and 1 (Barnard, 1989).

The problems discussed above, due to the discreteness of the distribution, have caused much controversy in the statistical literature, particularly with the analysis of data collected to compare two proportions. Following Lancaster (1952, 1961) the mid-P approach has been advocated more widely recently (e.g. Stone (1969), Plackett (1984), Miettinen (1985), Franck (1986), Rothman (1986), Williams (1988), Barnard (1989), Hirji (1991), Hirji *et al.* (1991) and Upton (1992)). Barnard (1989) recommended giving both the P- and the mid-P values, arguing that the former measures the statistical significance when the data under analysis are judged alone, whereas the latter is the measure of the strength of evidence against the hypothesis under test that it is appropriate to use when the evidence is combined with that from other studies. When the null hypothesis is true, the mean mid-P value is 0.5 and this property makes it particularly suitable in making an overall assessment by combining results of different studies, for instance by the probability integral transformation. Since it is rare that the results of a single study are used without support from other studies it seems reasonable to give more emphasis to the mid-P value.

3. Mid-P confidence intervals

The mid-P concept has generally been discussed in terms of significance testing but the concept can be extended to the calculation of confidence intervals in an estimation approach (Rothman, 1986; Mehta and Walsh, 1992). One definition of a confidence interval at, say, the 95% level is that it contains those values of θ that are not rejected by a significance test at the 5% level. An important distinction is that, whereas it is impossible to produce a satisfactory test at the 5% level for a predetermined θ_0, it is possible to define limits θ_L and θ_U as the values that are on the borderline of significance by one-sided tests at the 2.5% level. It is perfectly reasonable to specify the confidence coefficient in advance at some conventional value, such as 95%, whereas for a significance test it is desirable to estimate P as precisely as possible, given that it is impossible to construct a test at exactly the 5% level.

This approach gives mid-P 95% confidence limits as those values θ_L and θ_U which satisfy the equations

$$P_m(r, \theta_L) = \sum_{i=l}^{r} p_i(\theta_L) - \tfrac{1}{2}p_r(\theta_L) = 0.975, \tag{7}$$

$$P_m(r, \theta_U) = \sum_{i=l}^{r} p_i(\theta_U) - \tfrac{1}{2}p_r(\theta_U) = 0.025. \tag{8}$$

The mid-P limits are more tedious to calculate than the corresponding limits using the more conventional tail as they are not included in standard sets of tables and there is no direct formula such as those involving the F-distribution for a binomial probability (see Miettinen (1970)) or the χ^2-distribution for a Poisson parameter (see Liddell (1984)). The limits are available by using STATXACT (CYTEL Software Corporation, 1991) or may be obtained fairly readily by using a

personal computer by setting up the expression to be evaluated using a general argument, and then finding the values that give mid-P tails of 0.975 or 0.025 by trial and error.

4. Specific cases

Three simple specific cases will be considered but the method also applies in more complex situations such as the estimation of a common odds ratio for a 2×2 table stratified by a third variable (Mehta and Walsh, 1992).

4.1. *Binomial distribution*

The parameter to be estimated is the probability θ of a success in each of n independent trials. Then

$$p_i = \frac{n!}{i!\,(n-i)!}\theta^i(1-\theta)^{n-i}. \tag{9}$$

As an example consider the observation $r = 5$ from a binomial distribution with $n = 20$. The point estimate of θ is clearly 0.25. The mid-P limits of this estimate are given by

$$p_0 + p_1 + p_2 + p_3 + p_4 + \tfrac{1}{2}p_5 = 0.975 \text{ or } 0.025.$$

The exact 95% mid-P confidence limits were found to be 0.098 and 0.470.

4.2. *Poisson count*

The parameter to be estimated is the expectation μ. Then

$$p_i = \frac{\exp(-\mu)\mu^i}{i!}. \tag{10}$$

For example, suppose that a count from a Poisson distribution gave $r = 2$. Then the point estimate of μ is 2. The mid-P limits are given by

$$p_0 + p_1 + \tfrac{1}{2}p_2 = 0.975 \text{ or } 0.025,$$

i.e. $\mu_L = 0.335$ and $\mu_U = 6.61$. Cohen and Yang (1994) have presented a table for counts up to 100 giving the 90%, 95% and 99% mid-P confidence limits.

4.3. *2×2 table*

Suppose that the frequencies in a 2×2 table are as in Table 1. The association between rows and columns is frequently expressed in terms of the odds ratio ψ. Then the probability that $a = i$ is given by

$$P_i(\psi) = \frac{\binom{r_1}{i}\binom{r_2}{s_1 - i}\psi^i}{\sum_j \binom{r_1}{j}\binom{r_2}{s_1 - j}\psi^j}, \tag{11}$$

where

$$\binom{r_1}{i} = \frac{r_1!}{i!\,(r_1 - i)!}, \tag{12}$$

TABLE 1
2 × 2 table

Row	*Column*		*Totals*
	1	*2*	
1	*a*	*b*	r_1
2	*c*	*d*	r_2
Totals	s_1	s_2	*N*

the binomial coefficient, and the summation for j is over all possible values consistent with the margins (Fisher, 1935). When $\psi = 1$ this expression simplifies to

$$P_i = \frac{\binom{r_1}{i}\binom{r_2}{s_1 - i}}{\binom{N}{s_1}} = \frac{r_1!\, r_2!\, s_1!\, s_2!}{N!\, i!\, (r_1 - i)!\, (s_1 - i)!\, (r_2 - s_1 + i)!}, \tag{13}$$

the expression familiar in the calculation of Fisher's exact test of the null hypothesis.

Mid-*P* confidence limits for ψ are the solutions of

$$\sum_{i \leqslant a} P_i(\psi) - \tfrac{1}{2} P_a(\psi) = 0.025 \text{ and } 0.975 \tag{14}$$

(Rothman, 1986; Mehta and Walsh, 1992).

As an example consider the data in Table 2 discussed by Fisher (1935) and Thomas (1971). The point estimate of ψ is $(2 \times 3)/(10 \times 15) = 0.04$. As shown by Thomas (1971) the exact 95% confidence interval, defined in terms of lower and upper tails of 0.025, for ψ is 0.003317–0.3632. The mid-*P* interval can be calculated as 0.004825–0.2958, either by using STATXACT or by trial and error. In the latter case initial values for iteration can be obtained by using a normal-based approximation such as the method of Cornfield (1956) (see Section 5).

5. Approximate limits

In some cases approximate limits may be obtained by using a normal approximation to the discrete distribution. Where a normal approximation is adequate *P*-values and mid-*P* values correspond to test statistics calculated with and without the correction for continuity respectively. Correspondingly confidence intervals and mid-*P* confidence intervals can be based on normal approximations by using and ignoring the continuity correction respectively, e.g. the method given by Cornfield (1956) for estimation of the confidence limits for an odds ratio from a 2 × 2 table of

TABLE 2
Convictions of like-sex twins of criminals

	Convicted	*Not convicted*	*Totals*
Dizygotic	2	15	17
Monozygotic	10	3	13
	12	18	30

frequencies. An iterative method for deriving Cornfield's limits is illustrated by Fisher (1962) and Fisher and Yates (1963).

6. Discussion

In the analysis of a 2×2 table the exact test and, therefore, the χ^2-test with Yates's correction for continuity have been criticized on the grounds that they are conservative in that a result that is significant, at say the 5% level, will be found in less than 5% of hypothetical repeated random samples from a population in which the null hypothesis is true. This is a consequence of the discrete nature of the data and the problem is reduced by stating the precise level of P. Another source of criticism has been that the tests are conditional on the observed margins which frequently would not all be fixed. In the earlier example, we could imagine repetitions of sampling in which 17 dizygotic twins were compared with 13 monozygotic twins but in many of these samples the number of convicted twins would differ from 12. The conditional argument is that, whatever inference can be made about the association between twin type and conviction, it must be made within the context that exactly 12 twins were convicted. If this number had been different then the inference would have been made in this different context, but that is irrelevant to inferences that can be made when there are 12 convicted twins. Therefore, we do not accept the various arguments that have been put forward for rejecting the exact test based on consideration of possible samples with different totals in one of the margins. The issues were discussed by Yates (1984) and in the ensuing discussion, and by Barnard (1989) and Upton (1992), and we shall not pursue this point further. Clearly those who disagree with the conditional approach will reach different conclusions on the most appropriate way of calculating the significance level and producing a confidence interval (e.g. Rice (1988)).

Nevertheless the exact test and the corrected χ^2-test using the precise P-value are conservative since the average value of the significance level, when the null hypothesis is true, exceeds 0.5. The mid-P value has an average value of 0.5 and so is more appropriate particularly when combining results from several studies.

The corresponding mid-P confidence intervals do not necessarily have the specified coverage probabilities. Mehta and Walsh (1992) showed by simulation for a range of situations, where a common odds ratio was estimated over a stratified set of 2×2 tables, that the 95% mid-P confidence interval contained the true odds ratio in at least 95% of samples. Further extensive simulations have been described by Vollset (1993) for the binomial distribution and by Cohen and Yang (1994) for the Poisson distribution. In both these studies the coverage probability varies a little with the true value of the parameter but clusters satisfactorily around the nominal value except for extreme values of the parameter (e.g. the Poisson parameter near 0). At these extremes the non-coverage probability is approximately halved (for example the coverage becomes 97.5% rather than 95%). The reason for this is the need for one-sided limits when extreme observations are made, as indicated in the next paragraph. The results of Vollset (1993) and Cohen and Yang (1994) contrast with those of Mehta and Walsh (1992) who found the mid-P confidence interval to be conservative even though they used a different policy for dealing with extreme values that gives a narrower interval.

In this paper the 95% confidence interval has been defined such that each tail has a probability of 0.025. Narrower confidence intervals may be produced if this criterion of symmetry is not followed (Baptista and Pike, 1977) but this approach has the disadvantage that the separate limits have no direct meaning, i.e. the upper limit is not necessarily the value below which the confidence coefficient is 0.975. The distinction is most marked when the observed frequency is one of the extreme values. For example, suppose that a Poisson variable took the value 0. The point estimate of μ is 0 and the only possible value of the lower limit is 0. The probability that the lower limit exceeds the true value of μ is 0 instead of the nominal value of $2\frac{1}{2}\%$, and a possibility is to calculate the upper limit to correspond to a tail of 5% (Mehta *et al.*, 1985; CYTEL Software Corporation, 1991). This provides the narrowest 95% confidence interval but it seems preferable that the upper limit corresponds to $2\frac{1}{2}\%$ in the tail so that the interpretation of that limit is uniform for zero and non-zero

counts. It is impossible to find a lower limit with this interpretation, so that strictly this limit does not exist. This rationale is similar to that used in recommending that a two-sided significance level should be double the one-sided level. In situations where a one-sided significance test is considered appropriate then there corresponds a one-sided confidence interval, i.e. an interval unbounded on one side and with a limit on the other side at which the data are just significant using a one-sided test at the 5% level.

References

Baptista, J. and Pike, M. C. (1977) Algorithm AS 115: Exact two-sided confidence limits for the odds ratio in a 2 × 2 table. *Appl. Statist.*, **26**, 214–220.

Barnard, G. A. (1989) On alleged gains in power from lower P-values. *Statist. Med.*, **8**, 1469–1477.

Cohen, G. R. and Yang, S.-Y. (1994) Mid-P confidence intervals for the Poisson expectation. *Statist. Med.*, **13**, 2189–2203.

Cornfield, J. (1956) A statistical property arising from retrospective studies. In *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, vol. 4, pp. 135–148. Berkeley: University of California Press.

CYTEL Software Corporation (1991) *StatXact—Statistical Software for Exact Nonparameteric Inference, User Manual Version 2*. Cambridge: CYTEL Software Corporation.

Fisher, R. A. (1932) *Statistical Methods for Research Workers*, 5th edn. Edinburgh: Oliver and Boyd.

——— (1935) The logic of inductive inference. *J. R. Statist. Soc.* A, **98**, 39–54.

——— (1962) Confidence intervals for a cross-product ratio. *Aust. J. Statist.*, **4**, 41.

Fisher, R. A. and Yates, F. (1963) *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edn, pp. 6–7. Edinburgh: Oliver and Boyd.

Franck, W. E. (1986) P-values for discrete test statistics. *Biometr. J.*, **28**, 403–406.

Haber, M. (1986) A modified exact test for 2 × 2 contingency tables. *Biometr. J.*, **28**, 455–463.

Hirji, K. F. (1991) A comparison of exact, mid-P, and score tests for matched case–control studies. *Biometrics*, **47**, 487–496.

Hirji, K. F., Tan, S.-J. and Elashoff, R. M. (1991) A quasi-exact test for comparing two binomial proportions. *Statist. Med.*, **10**, 1137–1153.

Lancaster, H. O. (1949) The combination of probabilities arising from data in discrete distributions. *Biometrika*, **36**, 370–382.

——— (1952) Statistical control of counting experiments. *Biometrika*, **39**, 419–422.

——— (1961) Significance tests in discrete distributions. *J. Am. Statist. Ass.*, **56**, 223–234.

Liddell, F. D. K. (1984) Simple exact analysis of the standardized mortality ratio. *J. Epidem. Commty Hlth*, **38**, 85–88.

Mehta, C. R., Patel, N. R. and Gray, R. (1985) Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *J. Am. Statist. Ass.*, **80**, 969–973.

Mehta, C. R. and Walsh, S. J. (1992) Comparison of exact, mid-*p*, and Mantel–Haenszel confidence intervals for the common odds ratio across several 2 × 2 contingency tables. *Am. Statistn*, **46**, 146–150.

Miettinen, O. S. (1970) Estimation of relative risk from individually matched series. *Biometrics*, **26**, 75–86.

——— (1985) *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. New York: Wiley.

Plackett, R. L. (1984) Discussion on Tests of significance for 2 × 2 contingency tables (by F. Yates). *J. R. Statist. Soc.* A, **147**, 458.

Rice, W. R. (1988) A new probability model for determining exact *p*-values for contingency tables when comparing binomial proportions. *Biometrics*, **44**, 1–22.

Rothman, K. J. (1986) *Modern Epidemiology*. Boston: Little, Brown.

Stone, M. (1969) The role of significance testing: some data with a message. *Biometrika*, **56**, 485–493.

Thomas, D. G. (1971) Algorithm AS 36: Exact confidence limits for the odds ratio in a 2 × 2 table. *Appl. Statist.*, **20**, 105–110.

Tocher, K. D. (1950) Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika*, **37**, 130–144.

Upton, G. J. G. (1992) Fisher's exact test. *J. R. Statist. Soc.* A, **155**, 395–402.

Vollset, S. E. (1993) Confidence intervals for a binomial proportion. *Statist. Med.*, **12**, 809–824.

Williams, D. A. (1988) Tests for differences between several small proportions. *Appl. Statist.*, **37**, 421–434.

Yates, F. (1984) Tests of significance for 2 × 2 contingency tables (with discussion). *J. R. Statist. Soc.* A, **147**, 426–463.