Difference of Proportions from the Same Survey

Author(s): Alastair J. Scott and George A. F. Seber

Source: *The American Statistician*, Vol. 37, No. 4, Part 1 (Nov., 1983), pp. 319-320

Published by: American Statistical Association

Stable URL: <http://www.jstor.org/stable/2682774>

Accessed: 18/11/2009 12:39

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=astata>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*.

<http://www.jstor.org>

paired t test on the complete pairs (1–10) is $\bar{d} = -1.30$, $s_d^2 = 20.68$, $t = -.9040$ on 9 df.

Using the SAS GLM procedure we get a value of $\hat{\sigma}^2 = 93.05/9 = 10.3389 = \frac{1}{2}s_d^2$. The sum of squares for treatments is $8.45 = \frac{1}{2} \cdot 10 \cdot (-1.30)^2$. The F test is .8173, which is .9040².

The ANOVA table (Table 2b) is affected by the additional observations only in the source due to blocks (or pairs). With the phony observations included the sum of squares is 454.34. With them omitted it is 325.45.

CONCLUSION

From this analysis, it seems clear that excluding unpaired observations is the simplest method to use. No additional information about treatment differences is obtained by including these unpaired data. However, it seems wise (and statistically responsible) to describe the characteristics of the unpaired data in case there are biases in the missed data.

Table 2b. SAS Results

Source	df	SS	MS	F
<i>a. ANOVA (phonies omitted)</i>				
Pair	9	325.45	36.16	3.50
Treatment	1	8.45	8.45	.8173
Error	9	93.05	10.34	
$\hat{\beta} = -1.30$				
<i>b. ANOVA (phonies included)</i>				
Pair	12	454.34	37.86	3.66
Treatment	1	8.45	8.45	.82
Error	9	93.05	10.34	

REFERENCES

- ARMITAGE, P. (1971), *Statistical Methods in Medical Research*, New York: John Wiley.
 MORRISON, D.F. (1973), "A Test for Equality of Means With Missing Data on One Response," *Biometrika*, 60, 101–105.
 NETER, J., and WASSERMAN, W. (1974), *Applied Linear Statistical Models*, Homewood, Ill.: Richard D. Irwin.

Difference of Proportions From the Same Survey

ALASTAIR J. SCOTT and GEORGE A.F. SEBER*

The problem of estimating the difference between two proportions for the same sample survey occurs often in practice and should be considered in introductory statistics courses.

KEY WORDS: Sample survey; Difference of two proportions.

1. INTRODUCTION

Although all elementary statistics books deal with the problem of estimating the difference between two proportions from independent samples, no attention is given to the equally important problem of estimating the difference between two proportions for the same sample survey. This problem is well known to readers-at-large in the form of preelection opinion polls. It has been our experience that newspapers (and students) frequently get it wrong and, at best, assume that estimated proportions are statistically independent. Since this topic has such a high profile, we believe that it should play a more prominent role in introductory sta-

tistics courses and elementary statistics books. The methods are well known, of course, but the derivations involving underlying multinomial and multihypergeometric distributions are usually fairly complex. However, with care, it is possible to handle them in an elementary fashion using only the binomial distribution and the univariate central limit theorem (CLT).

2. MULTINOMIAL SAMPLING

Suppose we have a population of N people of which N_i ($i = 1, 2, 3$) have preference for candidate i . A random sample of size n is taken without replacement and X_i are found to have preference i . If $f = n/N$, the sampling fraction, is small, say less than .05, then we can approximate sampling without replacement by sampling with replacement, and the joint distribution of the X_i is approximately multinomial with probability function

$$g(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}, \quad 0 \leq x_i \leq n, \quad \sum_i x_i = n,$$

where $p_i = N_i/N$. This distribution has the property that X_1 , X_2 , and $X_1 + X_2$ each have binomial distributions with parameters p_1 , p_2 , and $p_1 + p_2$. Such properties follow by simply observing that each pair of the three outcomes for the N independent trials can be merged into a single outcome, thus converting the multinomial experiment into a binomial one. Thus it is straightfor-

*Alastair J. Scott and George A.F. Seber are Professors of Statistics in the Department of Mathematics & Statistics, Auckland University, Private Bag, Auckland, New Zealand.

ward to find expressions for the variances of X_1 , X_2 , and $X_1 + X_2$, and then find $\text{cov}(X_i, X_j)$ using the well-known trick

$$\begin{aligned}\text{cov}(X_i, X_j) &= \frac{1}{2}\{\text{var}(X_i + X_j) - \text{var}(X_i) - \text{var}(X_j)\} \\ &= -np_i p_j.\end{aligned}$$

Now $\hat{P}_i = X_i/n$ is an unbiased estimate of p_i , and $E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2$. Also

$$\begin{aligned}\text{var}(\hat{P}_1 - \hat{P}_2) &= \text{var}(\hat{P}_1) + \text{var}(\hat{P}_2) - 2\text{cov}(\hat{P}_1, \hat{P}_2) \\ &= (p_1 q_1 + p_2 q_2 + 2p_1 p_2)/n \quad (q_i = 1 - p_i) \\ &= [p_1 + p_2 - (p_1 - p_2)^2]/n.\end{aligned}$$

For n sufficiently large, say $\text{minimum}(n\hat{p}_i\hat{q}_i) > 10$, we can invoke the multivariate CLT, as applied to the multinomial distribution, and obtain an approximate 95% confidence interval

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \{[\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2]/n\}^{1/2}.$$

Alternatively, if we define a random variable Y_j that takes the value +1 if the j th sampled person prefers candidate 1, -1 for candidate 2, and 0 otherwise, then $\hat{P}_1 - \hat{P}_2 = \bar{Y}$ and we only need the univariate CLT.

As shown by Goodman (1965), the method can be extended to give Bonferroni simultaneous confidence intervals for k such differences from the same survey. We simply use the $100\alpha/(2k)\%$ point of the standard normal distribution, instead of 1.96, in the interval just given.

For planning purposes, an upper bound of the square root term is $n^{-1/2}$, which compares with $\frac{1}{2}n^{-1/2}$ (setting $p_i = \frac{1}{2}$) for a single estimate $\hat{p}_i$, and $(2n)^{-1/2}$ for the difference of independent sample proportions. The large sample statistic for testing $H_0: p_1 = p_2$ takes the particularly simple form $Z_0 = (\hat{P}_1 - \hat{P}_2)/\{(\hat{P}_1 + \hat{P}_2)/n\}^{1/2}$. The special case of $p_2 = 1 - p_1$ is also of interest.

3. HYPERGEOMETRIC SAMPLING

If the sampling fraction cannot be ignored, then the

joint distribution of the X_i is multihypergeometric, namely,

$$\begin{aligned}g(x_1, x_2, x_3) &= \binom{N_1}{x_1} \binom{N_2}{x_2} \binom{N_3}{x_3} / \binom{N}{n}, \\ 0 \leq x_i &\leq \min(n, N_i), \sum_i x_i = n.\end{aligned}$$

Using the same argument as in the preceding section, we see that X_1 , X_2 , and $X_1 + X_2$ have hypergeometric distributions. For example, the probability function of X_1 (and of $X_2 + X_3$) is given by

$$g(x_1) = \binom{N_1}{x_1} \binom{N_2 + N_3}{n - x_1} / \binom{N}{n}.$$

From the properties of this distribution we know that $\hat{P}_i$ is still unbiased and $\text{var}(\hat{P}_i) = p_i q_i (N - n)/n(N - 1)$. We see that the algebra for the multinomial distribution goes through for the multihypergeometric except that every term is multiplied by $(N - n)/(N - 1)$, or approximately $1 - f$.

4. DISCUSSION

The difference, $p_1 - p_2$, may not always be the most appropriate target for our inferences. Cochran (1977, Sec. 3.8) suggests estimating the conditional probability $p_1/(p_1 + p_2)$ instead and making inferences conditional on the observed total $x_1 + x_2$. However, the simple difference, $p_1 - p_2$, is so widely quoted in newspapers and other nontechnical sources that it seems worthwhile that students in elementary courses be able to associate a valid standard error with a quoted difference.

[Received December 1982. Revised June 1983.]

REFERENCES

- COCHRAN, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.
GOODMAN, L.A. (1965), "On Simultaneous Confidence Intervals for Multinomial Proportions," *Technometrics*, 7, 247-255.