BMC Medical Research Methodology

BioMed Central

Research article

Estimation of the correlation coefficient using the Bayesian Approach and its applications for epidemiologic research

Enrique F Schisterman*[1], Kirsten B Moysich[2], Lucinda J England[1] and Malla Rao[1]

Address: [1]Division of Epidemiology, Statistics and Prevention National Institute of Child Health and Human Development / National Institute of Health, Bethesda, MD, USA and [2]Department of Cancer Control, Epidemiology and Biostatistics, Roswell Park Cancer Institute, Buffalo, NY, USA

Email: Enrique F Schisterman* - schistee@mail.nih.gov; Kirsten B Moysich - Kirsten.Moysich@RoswellPark.org; Lucinda J England - englandl@mail.nih.gov; Malla Rao - MRao@niaid.nih.gov

* Corresponding author

Published: 25 March 2003

BMC Medical Research Methodology 2003, **3**:5

Received: 6 September 2002
Accepted: 25 March 2003

This article is available from: http://www.biomedcentral.com/1471-2288/3/5

Abstract

Background: The Bayesian approach is one alternative for estimating correlation coefficients in which knowledge from previous studies is incorporated to improve estimation. The purpose of this paper is to illustrate the utility of the Bayesian approach for estimating correlations using prior knowledge.

Methods: The use of the hyperbolic tangent transformation ($\rho = \tanh \xi$ and $r = \tanh z$) enables the investigator to take advantage of the conjugate properties of the normal distribution, which are expressed by combining correlation coefficients from different studies.

Conclusions: One of the strengths of the proposed method is that the calculations are simple but the accuracy is maintained. Like meta-analysis, it can be seen as a method to combine different correlations from different studies.

Background

The correlation coefficient is a standard measure of association between two random variables and is widely used in epidemiology. As such, considerable attention has been given to its interpretation [1–3] as well as to the methods for correcting attenuation due to random measurement error [4,5]. Strategies for correcting measurement error require knowledge about the reliability of the measurements [2] for the use of an alloyed gold standard [6] to estimate reliability coefficients. In many epidemiological studies, the reliability of the measurements is unknown making it impossible to correct for attenuation.

Classical methods are based solely on collected data, and ignore any prior knowledge of the association under investigation. The Bayesian approach is one alternative for estimating correlation coefficients in which knowledge from previous studies is incorporated to improve estimation. The purpose of this paper is to illustrate the utility of the Bayesian approach. The summarizing properties and correction for measurement error of the Bayesian approach will be demonstrated. To illustrate this method, the correlation between maternal weight gain during pregnancy and infant birth weight will be examined.

Statistical Methods

Bayes' Theorem holds that a prior state of knowledge offers relevant information for statistical analyses. To update beliefs about a hypothesis, Bayes' Theorem is used to calculate the posterior probability of the hypothesis, such as correlation coefficient ρ. As such, Bayes' Theorem [7] holds that the posterior probability ρ is given by the following formula:

$$P(\rho \mid data) = \frac{P(data \mid \rho) P(\rho)}{P(data)} \qquad (1.1)$$

The factor P(data|ρ) is the likelihood function evaluated at ρ or the data collected from the investigator's study. The P(ρ) depends upon information present before the study, i.e., prior probability. The term $1/P(data)$ should be viewed as a factor that makes the total probability equal to 1 when adding over all possible ρ's; that is, the denominator P(data) is the sum or integral of the numerator over all ρ's. It is often referred to as the normalizing constant. Bayes' Theorem [8] can be rewritten as such:

Posterior Probability ∝ Likelihood × Prior Probability (1.2)

where ∝ means proportional to.

We suppose that the two variables of interest, X and Y, follow a bivariate normal distribution with means μ_x and μ_y, variances σ_x and σ_y, respectively, and correlation coefficient ρ(x,y) = ρ. We will use the following conventional notation to represent the sample mean, variance and covariance:

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{y} = \frac{\sum y_i}{n}$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2$$

and

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}).$$

Also, as a reminder, the sample correlation coefficient *r* is defined by:

$$r = \frac{Sxy}{\sqrt{SxxSyy}} \qquad (1.3)$$

Using standard reference priors for μ_x, μ_y, σ_x, and σ_y, and applying (1.2), a reasonable approximation to the posterior density [4] of ρ is given by

$$P(\rho \mid x, y) \propto P(\rho) \frac{(1-\rho^2)^{(n-1)/2}}{(1-\rho \times r)^{n-(3/2)}} \qquad (1.4)$$

After making the substitution ρ = tanh ξ and *r* = tanh z, we find that ξ is approximately normal with mean z and variance 1/n. These results were derived in a series of complicated substitutions by Fisher [10] and are described in detail elsewhere [4].

One of the most important properties of the hyperbolic tangent transformation (ρ = tanh ξ and *r* = tanh z) is its capacity to take full advantage of the conjugate properties of the normal distribution, which is accomplished by combining correlation coefficients from different studies. As stated in (1.2), we need a prior and a likelihood function to find the posterior density, which will follow a normal distribution where:

$$\mu_{Posterior} = \varsigma^2_{pos} \times (n_{prior} \times \tanh^{-1} r_{prior} + n_{Likelihood} \times \tanh^{-1} r_{likelihood}) \qquad (1.5)$$

and variance:

$$\sigma^2_{Posterior} = \frac{1}{n_{prior} + n_{Likelihood}} \qquad (1.6)$$

In general, many different priors can be used in (1.4), but clearly the inference becomes easier if we choose a prior in the following form for c:

$$P(\rho) \propto (1 - \rho^2)^c \qquad (1.7)$$

The choice of c is an important one, since it will determine the weight the prior will have in estimation. If we do not have any information from previous studies, a common choice for c will be 0, that is, $p(\rho) \propto 1$. There are some other choices for c, such as -3/2 (referred to as the multiple parameter Jeffreys' rule) [8]. A detailed description of this concept is beyond the scope of this paper and is discussed elsewhere [9,10].

Application

Researchers have hypothesized that birth weight and maternal weight gain during pregnancy are correlated, especially in African American women. To examine this question, we utilized the data from the Angler Cohort Study.

Description of the Study Population

The New York State Angler Cohort Study was initiated in 1991 to characterize exposure to persistent toxic contaminants through the consumption of Lake Ontario sport fish in men and women of reproductive age. Potential relations between these exposures and various reproductive and developmental endpoints were also assessed. A description of the cohort and has been published elsewhere [11]. Briefly, the New York State Angler Cohort Study employed a cross-sectional design to survey a stratified random sample of men and women between the ages of 18 and 40 who bought fishing licenses in 16 upstate New York counties in close proximity to Lake Ontario. Detailed information has been complied for the children born to cohort members between 1986 and 1991 and includes data from birth certificates and maternal and newborn medical records. Of the 2430 women with singleton index births during the study time period, 2205 (91%) had both medical records and birth certificates available with no missing data relevant to the study question.

Among the index study group of children, the prevalence of low birth weight (<2500 grams) and pre-term delivery (<37 weeks) were 3.3 and 3.7 percent, respectively. The mean birth-weight was 3503 grams and the mean gestational age was 39.7 weeks. The majority of women were white (98.8%) and were married at the time of delivery (92.6%). For the current study, we restricted our analysis to African American women (n = 26). In these women, the mean weight gain was 29.61 ± 10.86 pounds, and the mean infant birth-weight was 3484 ± 462 grams.

Implementation of the Bayesian Methodology

The correlation between maternal weight gain during pregnancy and infant birth-weight in African American women was estimated using data from the Angler Cohort Study. It is known that maternal weight gain in this study was measured with error. The sample correlation coefficient between maternal weight gain and infant birth weight was r_{xy} = 0.27 (n = 26). This estimate differs greatly from that of a previous study (r_{xy} = 0.63) in which the maternal weight gain measurements were performed more precisely and were based in a large sample (n = 1026)[12].

We combined data collected in the Angler Cohort Study with information from the prior study using formulas (1.5) and (1.6). Specifically, we had a normally distributed prior and likelihood, which are conjugate functions. The posterior distribution then is normally distributed, with the following variance:

$$\sigma^2_{Posterior} = \frac{1}{n_{prior} + n_{Likelihood}} = \frac{1}{26 + 1026} = 0.0009$$

and mean

$\mu_{Posterior} = \varsigma^2_{\text{pos}} \times (n_{prior} \times \tanh^{-1} r_{prior} + n_{Likelihood} \times \tanh^{-1} r_{likelihood}) =$

$0.0009 \times (1026 \times \tanh^{-1} 0.63 + 26 \times \tanh^{-1} 0.266) = 0.691$

That is, Normal(Mean = 0.691, Variance = 0.0009), resulting in a point estimate of the correlation coefficient of tanh(0.691) = 0.598.

Since we know the posterior distribution, we also can calculate the 95% posterior probability interval, which is defined by

$$\mu_{Posterior} \pm 1.960 \times \sqrt{\left(\sigma^2_{Post}\right)} = 0.691 \pm 1.96 \times \left(0.0009\right)^{1/2}$$

that is, (0.63–0.75).

Using the hyperbolic tangent transformations, we obtained a corresponding interval for the posterior ρ (0.56–0.63). If we only based our conclusion on the collected data, the 95 percent confidence interval would be 0.27 ± 1.96 × $(1/26)^{1/2}$ or (-0.11 – 0.65). This corresponds to a 95% confidence interval for the correlation coefficient of (-0.11 – 0.57) in the original scale.

Discussion

The epidemiologic literature offers various methods for combining study results such as meta-analysis or correction procedures for measurement error concerns. Some of these approaches are not directly applicable to correlation studies while others are not practical due to the lack of suitable of statistical software and complicated mathematical formulations. In this paper, we introduce epidemiologists to an alternative method for estimating correlation coefficients, which is both simple and accurate.

In our example, the confidence interval of the correlation between maternal weight gain and infant birth weight based only on the data from the Angler Cohort Study was very wide and included zero. However, after applying Bayesian methods, the point estimate increased and the interval became very narrow. Assuming birth weight was perfectly measured, there are three potential explanations

for the marked differences in point and interval estimates: (1) there was sampling variation, (2) there was measurement error in weight gain measurements which attenuated the relation, or (3) the two samples came from two dissimilar populations. If the investigator suspects that the differences in point estimates and confidence intervals are due two sampling variation, small sample size, or random measurement error, the Bayesian approach provides a reasonable compromise.

Random measurement error attenuates correlation coefficients towards the null (i.e. toward no association). Strategies for correcting measurement error require knowledge about the reliability of the measurements [1–4], which is not usually available, or increasing the sample size, which is not usually possible. However, when there is knowledge of the correlation from previous studies, it can be coupled with the data collected and inference can be improved. Correlation estimates from previous studies can be used in this way to deattenuate the effects of measurement error. Furthermore, the Bayesian approach can be used to combine as many correlation coefficients as necessary to achieve better point estimates with narrower confidence intervals.

Bayesian methods have not received much attention in the biomedical literature, including epidemiology. The strength of the proposed method is that the calculations are simple. While more accurate approximations to this approach can be derived, the relative gains are small and are offset by the complexity of calculations [4]. Another noteworthy strength of the Bayesian method is that the confidence intervals can be interpreted as probabilities as they are based on a true probability function. This enables the investigator to assess the nature of the relation between two variables more intuitively.

The Bayesian approach can be used to correct for some attenuation due to measurement error and under-sampling of a referent population. However, we recognize that special attention should be given to the choice of prior when using Bayesian correction procedures, since differences in the correlation estimates between the sampled population and the prior may reflect population heterogeneity, and not sampling problems concerns, *per se*.

Conclusion

we encourage epidemiologists to consider the Bayesian approach as a tool for summarizing correlation coefficients across studies and for evaluating relations when measurement error and statistical power is of utmost concern. This approach is suitable for all sub-specialties of epidemiology.

Competing interests

None declared.

Authors' contributions

Each author contributed to the design, analysis and manuscript preparation.

Acknowledgment

We would like to thank Dr Buck for allowing us to use the Angler cohort data and for her insightful comments to previous versions of this paper.

References

1. Liu K **Measurement Error and its impact on partial correlation and multiple linear regression analyses** *Am J Epidemiol* 1988, **127:**864-874
2. Hakstian AR, Schroeder ML and Rogers WT **Inferential theory for partially disattenuated correlation coefficients** *Psychometrika* 1989, **54:**397-407
3. Millsap RE **Sampling variance in attenuated correlation coefficients: a Monte Carlo Study** *J Appl Psychol* 1988, **73:**316-319
4. Bashir SA and Duffy SW **The Correction of Risk Estimates for Measurement Error** *Ann Epidemiol* 1997, **7:**154-164
5. Dear KBG, Puterman ML and Dobson AJ **Estimating correlations from epidemiological data in the presence of measurement error** *Statistics in Medicine* 1997, **16:**2177-2189
6. Spiegelman D and Cassella M **Fully parametric and semi-parametric regression models for common events with covariant measurement error in main study/validation study designs** *Biometrics* 1997, **53:**395-409
7. Lee PM **Bayesian Statistics: An Introduction** *New York, Oxford University Press* 1989,
8. Berry DA and Stangl DK **Bayesian Biostatistics** *New York, Marcel Dekker* 1996,
9. Box GEP and Tao DR **Bayesian Inference in Statistical Analysis** *Reading, MA, Addison-Wesley* 1973,
10. Fisher RA **Frequency distributions of the values of the correlation coefficient in samples from an indefinitely large population** *Biometrika* 1915, **10:**507-521
11. Mendola P, Vena J and Buck GM **Exposure Characterization, Reproductive and Developmental Health in the New York Angler Cohort Study** *Great Lakes Research Review* 1995, **1:**2
12. Pickett KE, Abrams B and Selvin S **Maternal height, pregnancy weight gain, and birth weight** *Am J Hum Biol* 2000, **12:**682-687

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/5/prepub