

Student's s -statistic

David F. Andrews

Department of Statistics

University of Toronto

Toronto, Canada M5S 3G3

david@utstat.utoronto.ca

Student's t -statistic is a data-based linear transformation of an average. This paper proposes an extension i) to more general estimates including m.l.e.'s and M-estimates, and ii) to nonlinear transformations, so that the variance of the estimate is approximately constant. The expansion of the statistic and its properties are derived using basic procedures for the symbolic computation of asymptotic expansions.

Introduction

Some ninety years ago, Gosset (Student, 1908), introduced the t -statistic to calculate the probable error of the mean. His development was based on series expansions of moments and on Monte Carlo performed by shuffling 3000 pieces of cardboard. The statistic is a linear transformation of the mean designed to approximately standardize its first two moments.

Methods for estimates of more complex parameters have involved transformations to stabilize variance (Fisher, 1921), and distribution-free procedures such as the jackknife (Mosteller and Tukey, 1977) and the *BCA* bootstrap (Efron and Tibshirani 1992). Here we follow Gosset's approach and use estimates of moments to define closed-form confidence intervals based on an estimated variance-stabilizing transformation.

Methods for the symbolic computation of moments (Andrews and Stafford 1995) are used to define expressions for the quantities involved. For example, the full partition algorithm is used to compute expressions of the form $E[\bar{X}] = \mu_X$, $E[\bar{X} \bar{Y}] = \frac{1}{n}\mu_{XY} + (1 - \frac{1}{n})\mu_X\mu_Y$, etc. These may be estimated by approximating the moments by the corresponding averages of the data.

The constants (parameters) considered here are defined explicitly or implicitly in terms of expectations of the form: $E_F[\psi(\theta_F)] = 0$ where $\psi(\theta_F)$ is a function of the individual of the population that depends on the constant. These constants may be estimated from a sample by considering the sample to define an empirical distribution and defining the estimate $\hat{\theta}$ to satisfy $E_{F_n}[\psi(\theta_{F_n})] = \psi(\hat{\theta}) = 0$.

If the function $\psi(\hat{\theta})$ has a Taylor expansion about θ_F this expansion leads to the reversion of a series. This series expansion of the estimate may be computed simply using a quasi-Newton-Rapheson iteration beginning with $\theta_0 = \theta_F$ to yield a linear combination of products of expectations of ψ and its derivatives. The parameter is estimated with the same linear combination but of products of averages.

Example 1: Confidence intervals for the variance of a population

The proposed confidence intervals are introduced with the example of the population variance. The variance of the distribution F is defined by $\sigma^2 = E_F[X^2] - E_F[X]^2 = \mu_{X^2} - (\mu_X)^2$. It may be estimated by: $\widehat{\sigma^2} = E_{F_n}[X^2] - E_{F_n}[X]^2 = \bar{X^2} - (\bar{X})^2$. Since the estimate is expressed in terms of averages, *its* mean may be simply computed and estimated by

$$E_{F_n}[\widehat{\sigma}^2] = (1 - \frac{1}{n})(\overline{X^2} - (\overline{X})^2)$$

The variance of $\widehat{\sigma}^2$ may be similarly computed and estimated by

$$Var_{F_n}[\widehat{\sigma}^2] = n^{-1}(\overline{X^4} - 4\overline{X} \overline{X^3} - (\overline{X^2})^2 + 8(\overline{X})^2\overline{X^2} - 4(\overline{X})^4)$$

We consider a nonlinear transformation of the statistic $\widehat{\sigma}^2 - \sigma^2$:

$$\tilde{\tau} = \frac{(\exp\{B(\widehat{\sigma}^2 - \sigma^2)\} - 1)}{B} \approx (\widehat{\sigma}^2 - \sigma^2) + B\frac{(\widehat{\sigma}^2 - \sigma^2)^2}{2} + O_P(n^{-3/2})$$

This expression for $\tilde{\tau}$ involves products of $\widehat{\sigma}^2$ which is itself expressed in terms of products of averages. It follows that the moments of the transformed estimate may be computed mechanically. These lead, to expressions for estimates of the first two moments by $E_{F_n}[\tilde{\tau}]$ and $Var_{F_n}[\tilde{\tau}]$. Note that these expressions involve the transformation parameter B which has not yet been specified. The mean of the transformed statistic is computed to be

$$\frac{-2\mu_{X^2} - B\mu_{X^2}^2 + B\mu_{X^4}}{2n}.$$

The variance is computed to be

$$\frac{2B\mu_{X^2}^3}{n} + \frac{B^2\mu_{X^2}^4}{2n} - \frac{2B\mu_{X^3}^2}{n} + \left(1 - \frac{2}{n}\right)\mu_{X^4} - \frac{3B\mu_{X^2}\mu_{X^4}}{n} + \frac{B^2\mu_{X^4}^2}{2n} + \mu_{X^2}^2 \left(-1 + \frac{4}{n} - \frac{B^2\mu_{X^4}}{n}\right) + \frac{B\mu_{X^6}}{n}.$$

The estimates of these two moments are used to define limits for $\tilde{\tau}$ of the form

$$-E_{F_n}[\tilde{\tau}] \pm t_{\alpha/2, n-1} \sqrt{Var_{F_n}[\tilde{\tau}]}.$$

The resulting limits for $\tilde{\tau} = \widehat{\sigma}^2 - \sigma^2$ are back transformed and inverted to produce the confidence limits for σ^2

$$\widehat{\sigma}^2 + \log\{1 + B (- E_{F_n}[\tilde{\tau}] \pm t_{\alpha/2, n-1} \sqrt{Var_{F_n}[\tilde{\tau}]})\}/B.$$

It remains to compute the transformation parameter, B chosen to lessen the dependence of the variance of the estimate on the parameter.

The expression for the dependence of $Var[\widehat{\sigma}^2]$ on σ^2 is complex in general. One way to investigate this dependence is to consider a collection of distributions in a neighbourhood of F . For each distribution the variance of the estimate could be plotted against the parameter and the slope of this plot used to summarize one aspect of the dependence. This approach may be implemented by considering the empirical distributions F_n to form the neighbouring distributions in which case the slope of the plot, estimated by least-squares, is $\beta = Cov_F(\widehat{\sigma}^2, Var_{F_n}[\widehat{\sigma}^2])/Var_F[\widehat{\sigma}^2]$ The covariance is computed to be (to order $O(n^{-1})$):

$$\left(2 - \frac{28}{n}\right)\mu_{X^2}^3 + \left(-4 + \frac{10}{n}\right)\mu_{X^3}^2 + \left(-3 + \frac{26}{n}\right)\mu_{X^2}\mu_{X^4} + \left(1 - \frac{6}{n}\right)\mu_{X^6},$$

which can be estimated by approximating using averages. The resulting estimated slope $\hat{\beta}$ is used below to define the variance stabilizing transformation.

Consider now the simple transformation $\hat{\tau} = \exp\{B\widehat{\sigma}^2\}$. Taylor expansion about σ^2 yields $Var[\hat{\tau}] \approx \exp\{2B\sigma^2\}B^2Var[\widehat{\sigma}^2]$. Differentiating with respect to σ^2 and substituting $\partial Var[\widehat{\sigma}^2]/\partial \sigma^2 = \beta\sigma^2$ yields the change in variance with respect to σ^2 . Setting this change to 0 gives $B = -\beta/(2Var[\widehat{\sigma}^2])$, which when estimated using estimates of its components completes the definition of the desired intervals. The result is a closed-form expression for the confidence

Table 1

| Counts non-coverage above, A, and below, B, of 95% confidence intervals in 1000 trials | | | | | | | | | | | | | | | | |
|--|---------------------|----|----------|----|----------|----|----------|----|--------------------------|----|----------|---------|----------|----|----------|----|
| Parameter | Variance σ^2 | | | | | | | | Correlation $\rho = 0.8$ | | | | | | | |
| Distribution | Gaussian | | | | Uniform | | | | Gaussian | | | Uniform | | | | |
| Sample size | $n = 20$ | | $n = 40$ | | $n = 20$ | | $n = 40$ | | $n = 20$ | | $n = 40$ | | $n = 20$ | | $n = 40$ | |
| Tail | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A |
| χ^2 Fisher | 28 | 26 | 27 | 28 | 0 | 4 | 1 | 1 | 41 | 32 | 35 | 29 | 22 | 14 | 10 | 6 |
| BCA | 30 | 92 | 24 | 62 | 19 | 41 | 23 | 25 | 45 | 35 | 37 | 27 | 25 | 28 | 20 | 19 |
| s -interval | 21 | 89 | 20 | 60 | 3 | 44 | 16 | 24 | 25 | 36 | 28 | 26 | 25 | 32 | 17 | 19 |

limits expressed in terms of products of averages. This form may be implemented as a numerical function in, for example, S. The intervals are referred to here as s -intervals.

The coverages of the s -intervals are compared with those of the standard χ^2 intervals and the BCA bootstrap intervals based on 2000 bootstrap samples. 1000 simulations were generated for both samples sizes 20 and 40 from both the Gaussian and the uniform distribution. The counts of instances where the true value was below (B) and above (A) the interval are presented in Table 1. Since the intervals had nominal coverage of 95% these counts are expected to all be equal to 25 (with standard deviation approximately 5).

The s -intervals are similar to the bootstrap intervals although the former require the calculation of only 6 averages in contrast to the 2000 bootstrap variance estimates. Both are preferred to the exact intervals for the uniform distribution for $n > 40$.

Example 2: Confidence intervals for the correlation coefficient

The procedure used to define confidence intervals may be applied to the correlation coefficient $\rho = Cov[X, Y] / \sqrt{Var[X] Var[Y]}$. The only extension required is to first express the inverse roots in the definition of the statistic as a power series.

The coverages of the resulting intervals are compared in Table 1 with those based on Fisher's (1921) $(1/2)Log[(1+r)/(1-r)]$ transformation of the correlation coefficient r . This transformation may be derived from the differential equation relating the variance of r to the parameter, ρ , for the Gaussian distribution. This derivation, of course, raises the question as to what transformation should be used in the usual case in which the distribution is not known to be exactly Gaussian. Note, however, that Fisher's transformation could be justified simply on the need to remove the constraints $-1 \leq r \leq 1$.

Table 1 shows the close agreement between the s and the BCA bootstrap intervals. The former, however, are very efficiently computed from just a 19 averages. Both intervals are preferred those based on Fisher's transformation for the uniform distribution.

Example 3: Confidence intervals for logistic regression

Estimates in logistic regression are M.L.E's which have expansions in terms of products of averages. The expansion of the slope in a logistic regression was computed and the expansion of the confidence limits of the associated interval derived. These limits involved the calculation of

Table 2

| Counts non-coverage above, A, and below, B, of 95% confidence intervals for β in 1000 trials | | |
|---|----|----|
| Tail | B | A |
| Likelihood Ratio | 70 | 70 |
| s-interval | 61 | 14 |

only 9 averages. Table 2 presents the noncoverage of the s intervals and the standard likelihood ratio intervals for the true model $\text{logit}(P[Y|x]) = -0.5 + 3\log(x)$ when the logarithms were omitted in the fitted model: $\text{logit}(P[Y|x]) = \alpha + \beta x$ where $n = 200$ and $\log(x)$ is $U[-2, 2]$. Again the s -intervals are preferable to the standard likelihood ratio intervals. The Monte Carlo evaluation of BCA bootstrap intervals was not feasible because of the burden of computation.

Distribution-free confidence intervals have been constructed for parameters defined in terms of expectations. Symbolic computation has been used to compute closed-form expressions for the intervals. These involve the calculation of only a small number of averages. These intervals have coverage close to that of BCA bootstrap intervals but involve much less computation.

REFERENCES

- Andrews, D. F. and Stafford, J. E. (1995). Iterated full partitions. *Statistics and Computing*, **8**, 189-192.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman Hall, New York.
- Fisher, R. A. (1921). On the "probable error of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3-32.
- Mosteller, F. and Tukey, J. W. (1972). *Data Analysis and Regression*. Addison Wesley, Reading Mass.
- Student. (1908). The probable error of a mean. *Biometrika* **6**,1-25.

RESUMÉ

La statistique-t de Student est une transformation linéaire d'une moyenne, basée sur les données. Cet article propose une extension enfin de (i) obtenir des estimateurs plus générales y compris les estimateurs de vraisemblance et les M-estimateurs et (ii) utiliser des transformations nonlinéaire de sorte que la variance de l'estimateur soit approximativement constante. Les propriétés et l'expansion de la statistique sont dérivées par des méthodes de base pour le calcul symbolique d'expansions asymptotiques.