Generalized Variable Approach for Correlation and Weibull Analyses

Yanping Xia

Department of Mathematics
Southeast Missouri State University
Cape Girardeau, MO 63701

Joint work with Dr. Krishnamoorthy

July 31, 2009

Literature review

- Tsui, K. W. and Weerahandi, S.(1989): Generalized p-value in significance testing of hypotheses in the presence of nuisance parameters. *JASA* **84**, pp. 602–607.
- Weerahandi, S.(1993): Generalized confidence intervals, *JASA* **88**, pp. 899-905.
- Weerahandi S (1995) Exact statistical methods for data analysis. Springer: New York

Literature review

- Weerahandi S, Johnson RA (1992): Testing reliability in a stress-strength model when X and Y are normally distributed. *Technometrics* **34**, pp. 83–91.
- Zhou, L., Mathew, T. (1994): Some tests for variance components using generalized p-values, *Technometrics* **36**, pp. 394–402.
- Weerahandi, S. (1995): ANOVA under unequal error variances, *Biometrics* **51**, pp. 589–599.
- Weerahandi, S., and Berger, V. W. (1999): Exact inference for growth curves with intraclass correlation structure, *Biometrics* **55**, pp. 921–924.

Literature review

- Krishnamoorthy, K. and Lu, Y. (2003): Inferences on the common mean of several normal populations based on the generalized variable method, *Biometrics* **59**, pp. 237–247.
- Krishnamoorthy, K. and Mathew, T. (2004): One-Sided tolerance limits in balanced and unbalanced one-way random models based on generalized confidence limits, *Technometrics* **46**, pp. 44–52.
- Gamage, J., Mathew, T and Weerahandi, S. (2004): Generalized p-values and generalized confidence regions for the multivariate Behrens-Fisher problem, *Journal of Multivariate Analysis* **88**, pp. 177–189.
- Krishnamoorthy, K., Mukherjee,S. and Guo, Huizhen (2007): Inference on Reliability in Two-Parameter Exponential Stress-Strength Model, *Metrika* **65**, pp. 261–273

Generalized Pivotal Quantity (GPQ)

Let X and x be a random sample and an observed sample from a distribution $F(x;\theta,\delta)$, respectively, where θ is a scalar parameter of interest, and δ is a nuisance parameter.

A generalized pivotal quantity is a function of X, x, θ, and δ, denoted by $G(X;x,\theta,\delta)$, which satisfies the following conditions:

(i) For a given sample x, the distribution of $G(X;x,\theta,\delta)$ is free of unknown parameters.

(ii) $G(x;x,\theta,\delta)=\theta$.

Generalized Confidence Interval

Let G_p is the p-th quantile of $G(X; x, \theta, \delta)$, i.e.,

$$P\Big(G(X; x, \theta, \delta) \leq G_p\Big) = p.$$

For a given sample and $0 < \alpha < 1$, a $(1 - \alpha)$ generalized confidence interval for θ is given by

$$\left(G_{\frac{\alpha}{2}}, G_{1-\frac{\alpha}{2}}\right)$$

Generalized Test Variable (GTV)

The generalized test variable is of the form $T(X; x, \theta, \delta)$ which satisfies the following three requirements:

(i) $T(x; x, \theta, \delta)$ is free of any unknown parameters.

(ii) For a given sample x, the distribution of $T(X; x, \theta, \delta)$ is only dependent on θ.

(iii) For a given x and δ, the distribution of $T(X; x, \theta, \delta)$ is stochastically monotone in θ (i.e., stochastically increasing or decreasing in θ).

Generalized Test Variable (GTV)

The generalized test variable is of the form $T(X; x, \theta, \delta)$ which satisfies the following three requirements:

(i) $T(x; x, \theta, \delta)$ is free of any unknown parameters.

(ii) For a given sample x, the distribution of $T(X; x, \theta, \delta)$ is only dependent on θ.

(iii) For a given x and δ, the distribution of $T(X; x, \theta, \delta)$ is stochastically monotone in θ (i.e., stochastically increasing or decreasing in θ).

Using the GPQ, we can form the generalized test variable as follows:

$$T(X; x, \theta, \delta) = G(X; x, \theta, \delta) - \theta.$$

Functions of parameters

A function of generalized pivotal quantities is the generalized pivotal quantity for the corresponding function of the parameters:

$$G_{f(\theta_1,\dots,\theta_k)} = f\left(G_{\theta_1},\dots,G_{\theta_k}\right).$$

Furthermore, the corresponding GTV is given by

$$T_{f(\theta_1,\dots,\theta_k)} = f\left(G_{\theta_1},\dots,G_{\theta_k}\right) - f(\theta_1,\dots,\theta_k).$$

Notations

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Define

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \quad \text{and} \quad \mathbf{S} = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'.$$

Let

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \ldots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \ldots & \rho_{3p} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \ldots & \rho_{4p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \rho_{p4} & \ldots & 1 \end{pmatrix}$$

be the correlation matrix based on $\boldsymbol{\Sigma} = (\sigma_{ij})$, where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

A GPQ of ρ_{ij} based on a sample correlation matrix

Let $\boldsymbol{t}_r$ be the Cholesky factor of an observed sample correlation matrix:

$$\boldsymbol{r} = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & 1 & r_{23} & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \dots & 1 \end{pmatrix}$$

A GPQ for ρ_{ij} is given by

$$G_{\rho_{ij}} = \frac{\sum_{k=1}^{j} b_{ik} b_{jk}}{\sqrt{\sum_{k=1}^{i} b_{ik}^2}\sqrt{\sum_{k=1}^{j} b_{jk}^2}} \quad \text{for } i \geq j,$$

where $\boldsymbol{B} = \boldsymbol{t}_r \boldsymbol{V}^{-1} = (b_{ij})$ and $\boldsymbol{V} = (V_{ij})$ is a lower triangular matrix and it is well-known that the elements of $\boldsymbol{V}$ are independent with known distributions:

$$V_{ii}^2 \sim \chi_{n-i}^2, \ i = 1, \dots, p \ \text{ and } \ V_{ij} \sim N(0, 1), \ i > j.$$

Applications

We shall consider inferential procedures for the following problems.

1. Inference on a simple correlation coefficient (that is $p = 2$).
2. Comparing correlations of two independent bivariate normal populations.
3. Comparing two overlapping dependent correlations ($p \geq 3$).
4. Comparing two non-overlapping dependent correlations ($p \geq 4$).

Accuracy study

95% Upper Limits for ρ: $n = 3$

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
Exact	.906	.911	.922	.932	.942	.952	.961	.970	.980	.990	.995
GV	.906	.911	.922	.932	.942	.952	.961	.970	.980	.990	.995

The "exact" values of upper limits are computed by using the known pdf of the sample correlation (Anderson, Section 4.2, 1984). For our method (GV), we set $M = 1,000,000$ for Monte Carlo simulations.

K. Krishnamoorthy and Yanping Xia (2007): Inferences on correlation coefficients: one-sample, independent and correlated cases, Journal of Statistical Planning and Inference, 137, pp. 2362–2379.

MLEs

Let $x_1, \ldots, x_n$ be a sample from a Weibull(b, c) distribution with the pdf

$$f(x|b,c) = \frac{c}{b}\left(\frac{x}{b}\right)^{c-1} \exp\left\{-\left[\frac{x}{b}\right]^c\right\},\ x > 0,\ b > 0,\ c > 0.$$

The MLEs for the complete and censored cases can be obtained from Cohen (1965). For the complete case, the MLE $\widehat{c}$ of c is the solution to the equation

$$\frac{1}{\widehat{c}} - \frac{\sum_{i=1}^{n} x_i^{\widehat{c}} \ln(x_i)}{\sum_{i=1}^{n} x_i^{\widehat{c}}} + \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) = 0,$$

and the MLE $\widehat{b}$ of b is given by

$$\widehat{b} = \left(\sum_{i=1}^{n} x_i^{\widehat{c}}/n\right)^{1/\widehat{c}}.$$

GPQS for Weibull Parameters

Let $\widehat{b}_0$ and $\widehat{c}_0$ be the observed values of the MLEs based on a sample of n observations from a Weibull$(b,\ c)$ distribution. A GPQ for the scale parameter b is given by

$$G_b = \left(\frac{b}{\widehat{b}}\right)^{\frac{\widehat{c}}{\widehat{c}_0}} \widehat{b}_0 = \left(\frac{1}{\widehat{b}^*}\right)^{\frac{\widehat{c}^*}{\widehat{c}_0}} \widehat{b}_0.$$

A GPQ for c can be obtained as

$$G_c = \frac{c}{\widehat{c}}\widehat{c}_0 = \frac{\widehat{c}_0}{\widehat{c}^*} \quad ,$$

where

$$\widehat{c}/c \sim \widehat{c}^*, \quad \widehat{c}\ln\left(\widehat{b}/b\right) \sim \widehat{c}^*\ln\left(\widehat{b}^*\right).$$

Thoman, Bain,and Antle(1969).

Generalized Confidence Intervals for a Weibull Mean

Recall that the mean of a Weibull(b, c) distribution is given by

$$\eta = b\Gamma(1 + 1/c).$$

Thus, a GPQ for the mean η is given by

$$G_\eta = G_b\Gamma(1 + 1/G_c) = \left(\frac{1}{\widehat{b}^*}\right)^{\frac{\widehat{c}^*}{\widehat{c}_0}} \widehat{b}_0\Gamma\left(1 + \widehat{c}^*/\widehat{c}_0\right),$$

where $\widehat{b}_0$ and $\widehat{c}_0$ are observed values of $\widehat{b}$ and $\widehat{c}$ respectively.

Then the percentiles of G_η can be estimated using Monte Carlo simulation. Appropriate percentiles of G_η form a confidence interval for η. For example, the lower 5th percentile and the upper 5th percentile of G_η form a 90% CI for η.

Coverage study

Table: *Coverage Probabilities of 95% One-Sided Confidence Limits for a Weibull Mean*

$b = 1$

		lower limit					upper limit				
		c					c				
n	method	.5	1	2	3.5	5	.5	1	2	3.5	5
10	GV	.94	.95	.94	.94	.95	.95	.95	.94	.94	.95
	Wald	.98	.95	.93	.91	.91	.86	.89	.93	.93	.95
15	GV	.95	.94	.95	.94	.94	.95	.96	.95	.96	.96
	Wald	.98	.95	.93	.92	.93	.87	.90	.94	.95	.95
35	GV	.95	.95	.95	.95	.95	.96	.95	.95	.95	.95
	Wald	.97	.96	.94	.94	.93	.91	.93	.94	.95	.95
50	GV	.95	.95	.95	.95	.95	.95	.95	.95	.95	.95
	Wald	.97	.96	.94	.94	.93	.92	.93	.95	.95	.95

Wald: Asymptotic method in Colosimo and Ho (1999)

A GPQ for a Single Future Observation y

Let $x_1, ..., x_n$ be a sample from a Weibull(b, c) distribution. Then we can regard $y_i = \ln(x_i)$ as a sample from an extreme-value distribution with the location parameter $\mu = \ln(b)$ and the scale parameter $\sigma = 1/c$.

We have known that

(1) The MLEs of μ and σ are given by $\widehat{\mu} = \ln(\widehat{b})$ and $\widehat{\sigma} = 1/\widehat{c}$.

(2) $(y_i - \widehat{\mu})/\widehat{\sigma}$, $i = 1, ..., n$ are ancillary statistics and their distribution does not depend on μ or σ (Lawless 2003, p. 568).

A GPQ for y can be constructed as

$$G_y = \widehat{\mu}_0 + \frac{y - \widehat{\mu}}{\widehat{\sigma}}\widehat{\sigma}_0,$$

where $(\widehat{\mu}_0, \widehat{\sigma}_0)$ is an observed value of $(\widehat{\mu}, \widehat{\sigma})$.

Prediction Limits for y

Let $q^* = \frac{y^* - \widehat{\mu}^*}{\widehat{\sigma}^*}$. Then

$$G_y \sim \widehat{\mu}_0 + q^* \widehat{\sigma}_0.$$

Since the p-th quantile q_p^* of q^* can be computed by Monte Carlo simulation, a $(1-\alpha)$ generalized prediction interval for a single future observation from an extreme-value(μ, σ) can be obtained by

$$(\widehat{\mu}_0 + q^*_{\alpha/2}\widehat{\sigma}_0,\ \widehat{\mu}_0 + q^*_{1-\alpha/2}\widehat{\sigma}_0).$$

A $(1-\alpha)$ upper prediction limit (UPL) is given by

$$\widehat{\mu}_0 + q^*_{1-\alpha}\widehat{\sigma}_0.$$

UPL for Future Observations

A $(1-\alpha)$ UPL that will include at least l of m observations from each of r locations for an extreme-value(μ, σ) is given by

$$\widehat{\mu} + q_{n,r,m,l}\widehat{\sigma},$$

and for a Weibull(b, c) is given by

$$\exp\left(\widehat{\mu} + q_{n,r,m,l}\widehat{\sigma}\right).$$

where $q_{n,r,m,l}$ is the factor to be determined so that the coverage probability of the UPL is $(1-\alpha)$.

An Example

Table: *Vinyl Chloride Data from Clean Upgradient Ground-Water Monitoring Wells in ($\mu g/L$)*

5.1	2.4	0.4	0.5	2.5	0.1	6.8	1.2	0.5	0.6
5.3	2.3	1.8	1.2	1.3	1.1	0.9	3.2	1.0	0.9
0.4	0.6	8.0	0.4	2.7	0.2	2.0	0.2	0.5	0.8
2.0	2.9	0.1	4.0						

Bhaumik and Gibbons (2006)

An Example

Table: *95% Upper Prediction Limits for the Vinyl Chloride Data*

$n = 34$, $\widehat{\mu} = 0.6335$ and $\widehat{\sigma} = 0.990$

MLEs $\widehat{\mu} = \ln(\widehat{b}) = 0.6335$ and $\widehat{\sigma} = 1/\widehat{c} = 0.990$.

r	m	l	$u_{n,r,m,l}$	KLX	KMM	BG
1	2	1	0.461	2.974	2.893	2.931
10	2	1	1.079	5.483	5.203	5.224
10	3	1	0.659	3.618	3.479	3.521
10	3	2	1.296	6.797	6.369	6.330

BG Bhaumik, D. K., Gibbons, R. D. (2006). One-sided approximate prediction intervals for at least p of m observations from a gamma population at each of r locations. Technometrics 48, pp. 112–119.

KMM Krishnamoorthy, K., Mathew, T., Mukherjee, S. (2007). Normal based methods for a gamma distribution: prediction and tolerance intervals and stress-strength reliability. Technometrics 50, pp. 69–78.

KLX K. Krishnamoorthy, Yin Lin and Yanping Xia (2009). Confidence limits and prediction limits for a Weibull distribution based on the generalized variable approach. Journal of Statistical Planning and Inference 139, pp. 2675–2684.

Conclusions

- We have proposed the GV approach for correlation and Weibull analyses.
- Numerical results show that the GV approach is compatible with asymptotic methods and gives very satisfactory results even for small samples.
- The GV approach provides a unified treatment to all correlation problems considered.
- The GV approach can also be employed for statistical inferences on any functions of a normal covariance matrix, e.g., its determinant, trace, and eigenvalues.
- In addition, the GV approach can be used to make inferences for problems such as setting confidence limits on quantiles, survival probabilities, and estimating the ratio of two survival probabilities for Weibull distribution.

T

Th

Tha

Than

Thank

Thank y

Thank yo

Thank you

Thank you v

Thank you ve

Thank you ver

Thank you very

Thank you very much!

Thank you very much!

Thank you very much!