

[Page One](#)[Campus
Computing
News](#)[Winter Break
Hours](#)[Forget about
Santa Claus;
'GAUSS' is
coming to town!](#)[EagleMail News](#)[Lab-of-the-
Month: SLIS](#)[IDE RAID
Technology](#)[Today's Cartoon](#)

RSS Matters

[SAS Corner](#)[The Network
Connection](#)[List of the Month](#)[WWW@UNT.EDU](#)[Short Courses](#)[IRC News](#)[Staff Activities](#)[Subscribe to
Benchmarks
Online](#)

Research and Statistical Support

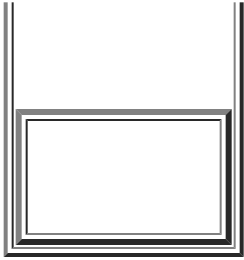
University of North Texas

RSS Matters

Dealing with Outliers in Bivariate Data: Robust Correlation

By [Dr. Rich Herrington](#), Research and Statistical Support Consultant

Last [time](#) we examined the smoothed bootstrap, this month we demonstrate the calculation of a robust correlation measure. The GNU S language, "R" is used to implement this procedure. R is a statistical programming environment that is a clone of the S and S-Plus language developed at Lucent Technologies. In the following document we illustrate the use of a GNU Web interface to the R engine on the "rss" server, <http://rss.acs.unt.edu/cgi-bin/R/Rprog>. This GNU Web interface is a derivative of the "Rcgi" Perl scripts available for download from the CRAN Website, <http://www.cran.r-project.org> (the main "R" Website). Scripts can be submitted interactively, edited, and be re-submitted with changed parameters by selecting the hypertext link buttons that appear below the figures. For example, clicking the "Run Program" button below creates 100 random numbers from a bivariate normal distribution; uses linear regression to fit a least squares line to two of the columns of data; and calculates person's product moment correlation. To view any text output, scroll to the bottom of the browser window. To view the scatterplot, select the "Display Graphic" link. The script can be edited and resubmitted by changing the script in the form window and then selecting "Run the R Program". Selecting the browser "back page" button will return the reader to this document.



Pearson's Product Moment Correlation

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from a bivariate distribution. The sample estimate of the population correlation ρ is given by:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

If X and Y are independent, $\rho = 0$. A test of $H_0: \rho = 0$ is based on the test statistic:

$$T = r \sqrt{\frac{n-2}{1-r^2}}$$

where n is the number of X,Y pairs of scores. If H_0 is true, T has a Student's t distribution with df = n-2 degrees of freedom if at least one of the marginal distributions is normal. Reject $H_0: \rho = 0$ if $|T| > t_{1-\alpha/2}$, for the $1-\alpha/2$ quantile of Student's t distribution with n-2

degrees of freedom. When the null hypothesis is true, the distribution of sample correlation coefficients tends to be normally distributed for increasing sample size. The standard error of this distribution of sample correlations is approximately:

$$\sigma_r = \frac{1}{\sqrt{N}}$$

When the sample size is reasonably large ($N \geq 50$), then it is possible to test the significance of the sample correlation coefficient by forming the usual z statistic and referring it to the normal distribution. In situations where one wishes to test for a non-zero value of the sample correlation coefficient, many sources have recommended Fisher's r-to-Z transformation when computing confidence intervals, but it is not asymptotically correct when sampling from nonnormal distributions. Moreover, simulation studies do not support Fisher's r-to-Z transformation for small sample sizes. The S script below samples from a bivariate normal distribution whose population correlation is .40. A scatterplot is plotted with a best fit regression line added to the scatterplot. A classical t-test of the correlation coefficient is performed, and the standard error under the null hypothesis is calculated. Change the script below to have different sample sizes and see the effect of sample size on the standard error; also see how the sample correlation fluctuates from sample to sample around the true population correlation value. With smaller sample sizes, the sample correlation coefficient can be quite different from the population value.

Robust Correlation: The Biweight Midcorrelation

For an introduction to robust statistics and robust measures of location, see the [July](#) issue of Benchmarks. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from a bivariate distribution. Following Wilcox (1997, page 197), let:

$$U_i = \frac{X_i - M_x}{9 \cdot MAD_x}$$

where M_x is the median, and MAD_x is the median absolute deviation. A sample estimate of the median absolute deviation is given by:

$$MAD_x = \text{MEDIAN} \left\{ \left| X_1 - M_x \right|, \dots, \left| X_n - M_x \right| \right\}$$

where M_x is the sample median. MAD has a finite sample breakdown point of approximately .5. Let V_i be a function of the Y_i scores as is U_i is for the X_i scores. The sample biweight midcovariance between X and Y is given by:

$$s_{bxy} = \frac{n \sum a_i (X_i - M_x) (1 - U_i^2)^2 b_i (Y_i - M_y) (1 - V_i^2)^2}{(\sum a_i (1 - U_i^2) (1 - 5U_i^2)) (\sum b_i (1 - V_i^2) (1 - 5V_i^2))}$$

where,

$$a_i = 1 \text{ if } -1 \leq U_i \leq 1, \text{ otherwise } a_i = 0$$

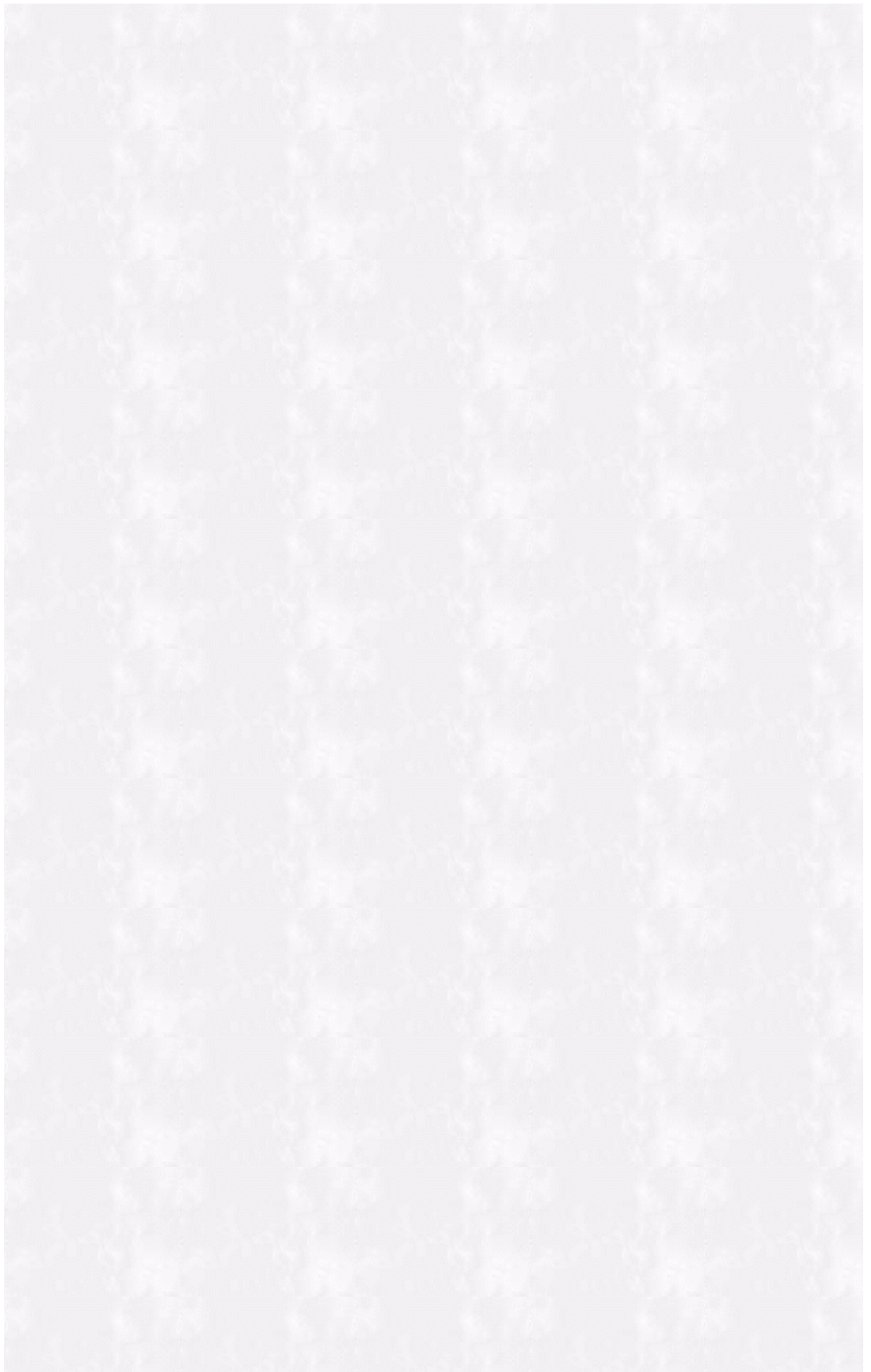
and

$$b_i = 1 \text{ if } -1 \leq V_i \leq 1, \text{ otherwise } b_i = 0.$$

An estimate of the biweight midcorrelation between X and Y is given by:

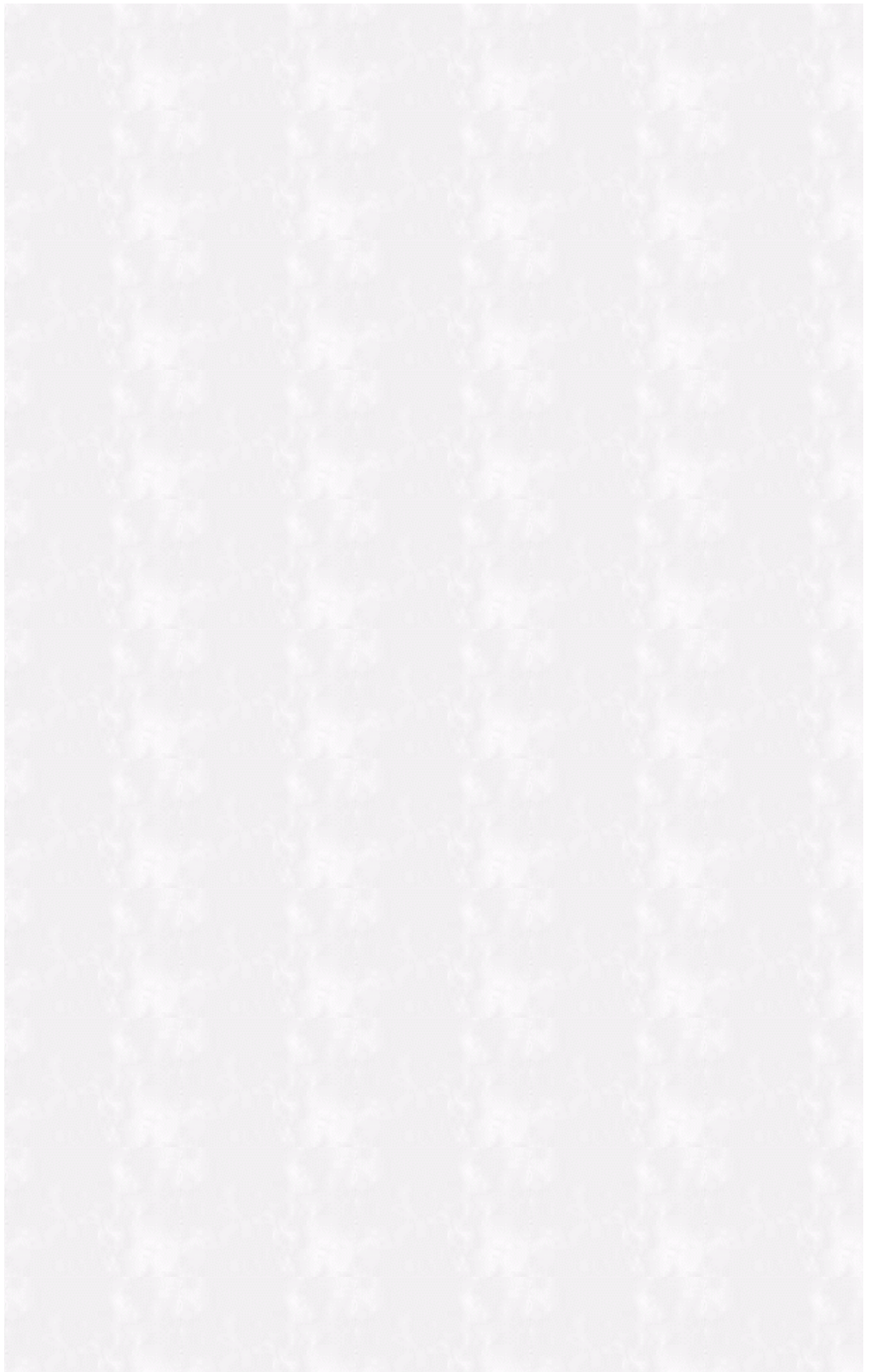
$$r_b = \frac{s_{bxy}}{\sqrt{s_{bxx} s_{byy}}}$$

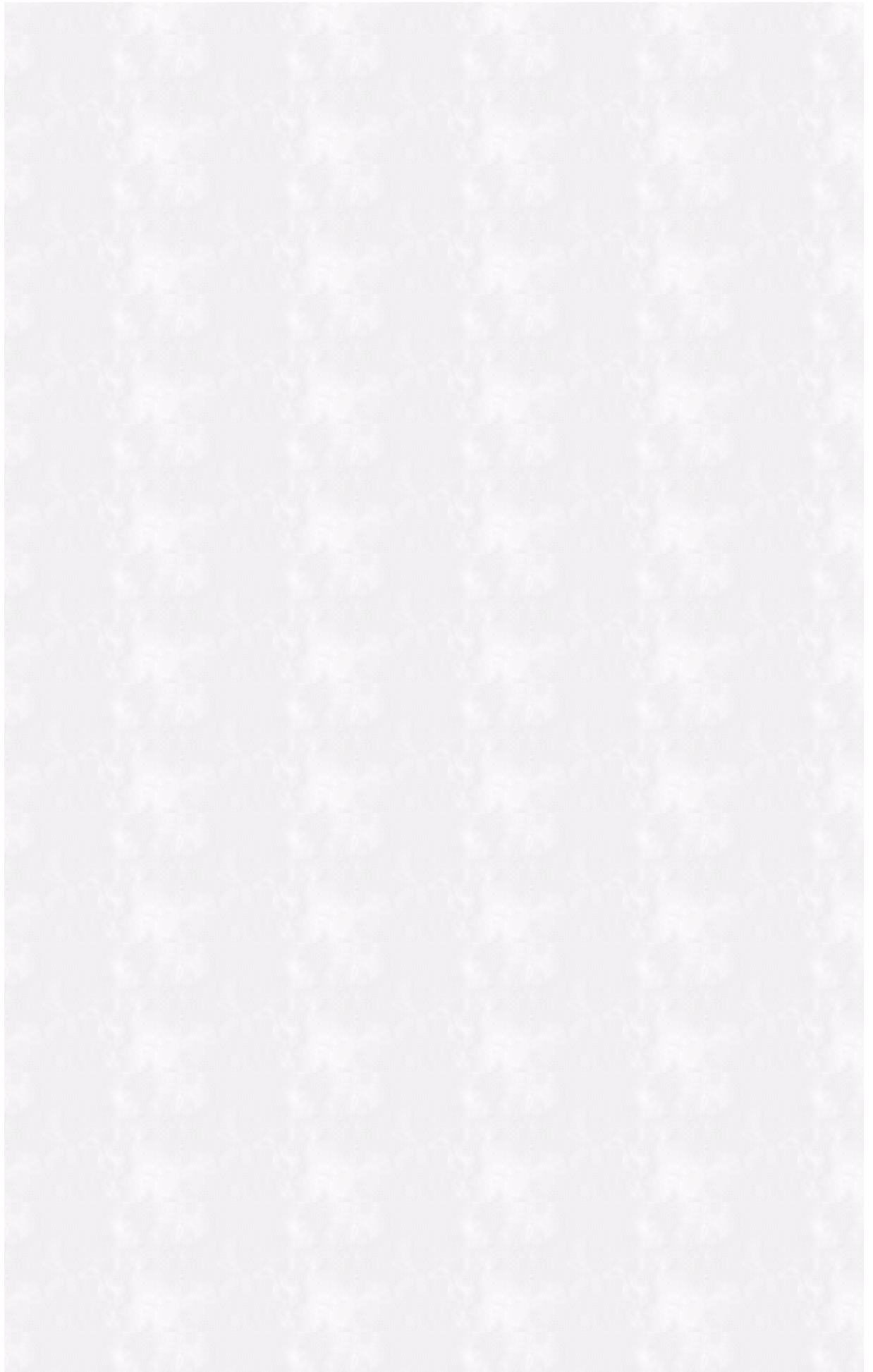
where s_{bxx} and s_{byy} are the biweight midvariance for the X and Y scores. The S script below calculates both the Pearson product-moment correlation and the biweight midcorrelation for a bivariate normal distribution whose population correlation is .40. An outlier is added to the sample, and then the two correlation coefficients are recalculated and compared again. A scatterplot with a best fit line is plotted after the outlier is added. Try changing various parameters and resubmitting the S script.

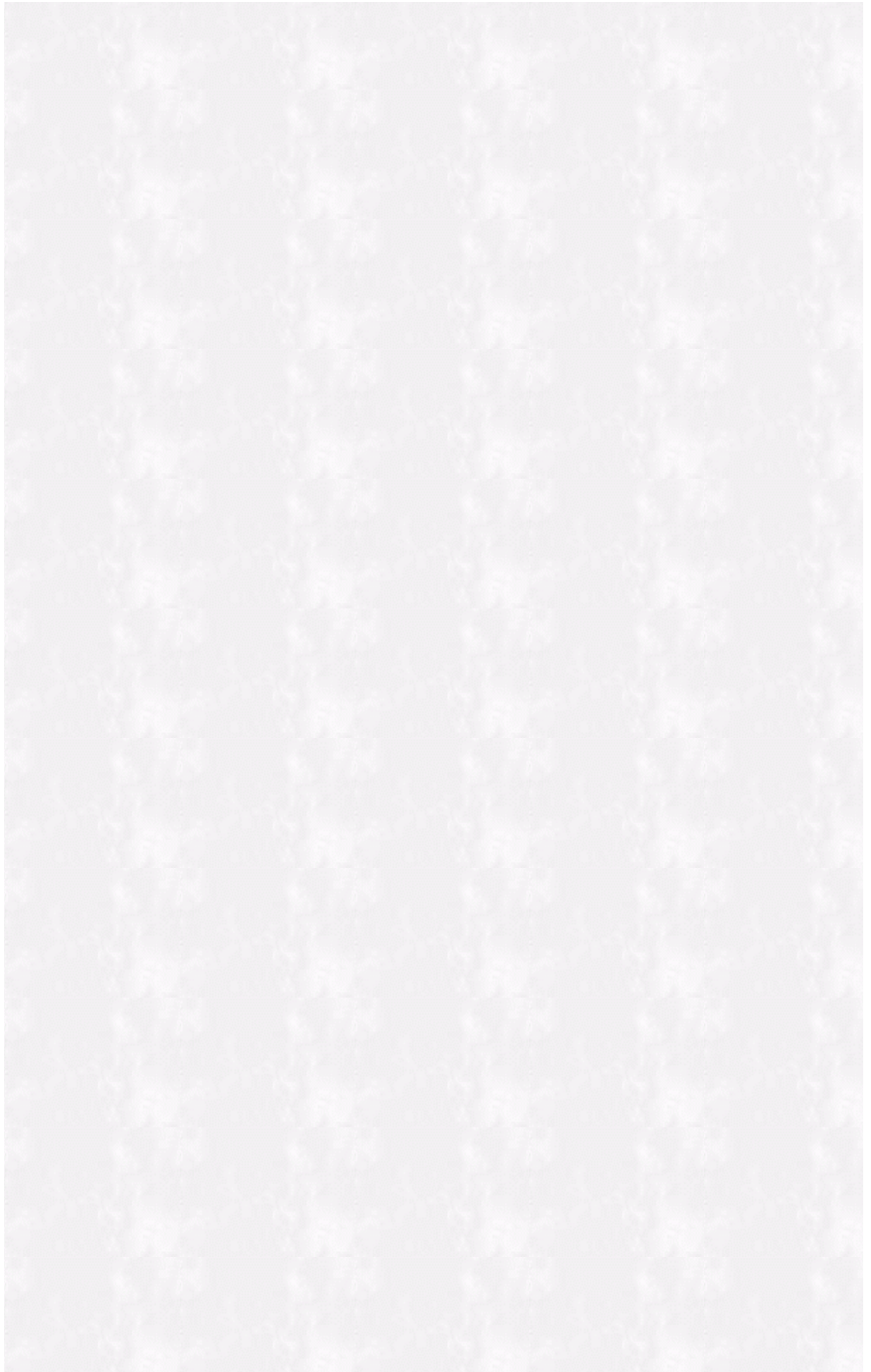


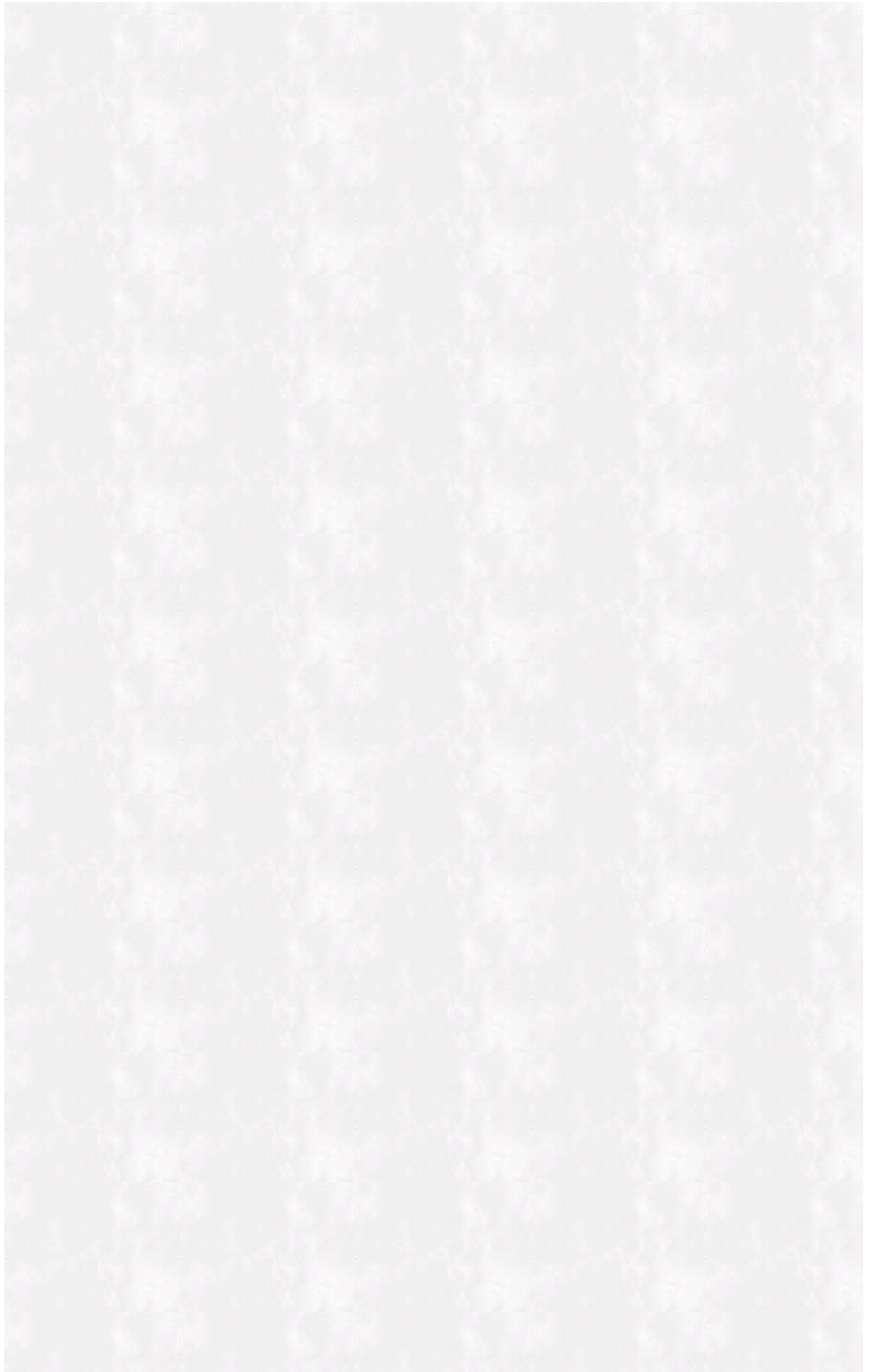
Calculating Empirical P-values, Empirical Power, and Confidence Intervals for a Robust Correlation Coefficient

The S script below calculates observed p-values, power, and confidence intervals for the biweight midcorrelation coefficient. A comparison is made between the Pearson correlation coefficient and the biweight midcorrelation before and after an outlier is added. Confidence intervals are calculated by simulating the null sampling distribution, and also by sampling from the alternate sampling distribution (Hall & Wilson, 1991). For details on calculating the percentile bootstrap, and calculating power using the percentile bootstrap, see the [September](#) issue of Benchmarks. For details on the percentile bootstrap see the [August](#) issue of Benchmarks. After running the following S script, try changing various parameters: sample size, population correlation, correlation coefficient (assign "est" to have value "bico" or "cor") and the size of the outlier (change the multiplying factor 2.5 to a smaller value). Changing "est" to have value "cor" will allow one to calculate the p-value and power of the observed sample for the Pearson product-moment correlation.









Results

The output from the S script above are listed below.

```
> ##### Report Results
>
> # Pearson Correlation and bi-weight midcorrelation
> # before outlier is added
>
> cor.nolie
[1] 0.4094627
> cor.nolie.test

      Pearson's product-moment correlation

data:  z[, 1] and z[, 2]
t = 2.7667, df = 38, p-value = 0.008698
alternative hypothesis: true correlation is not equal to 0
sample estimates:
      cor
0.4094627

> bicor.nolie
[1] 0.4222197
```

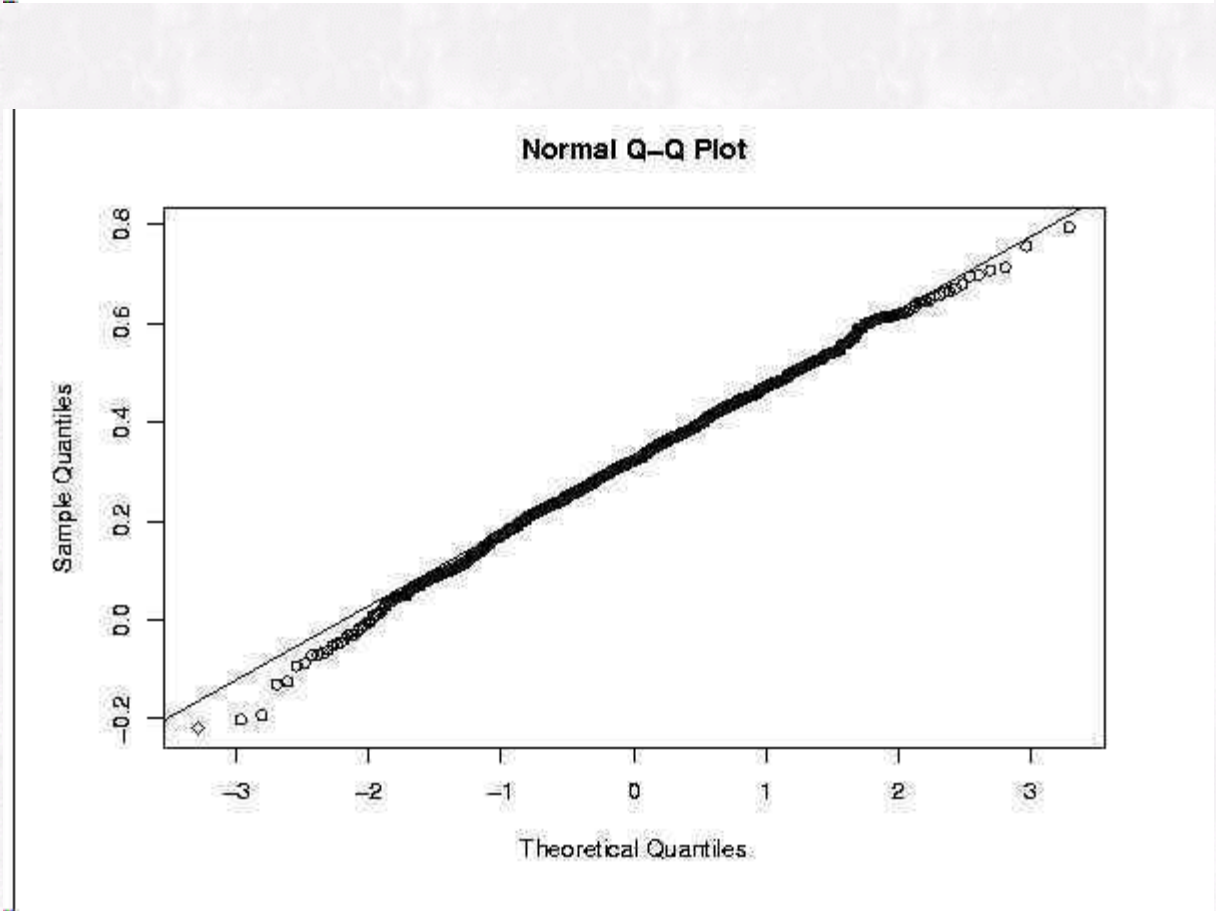
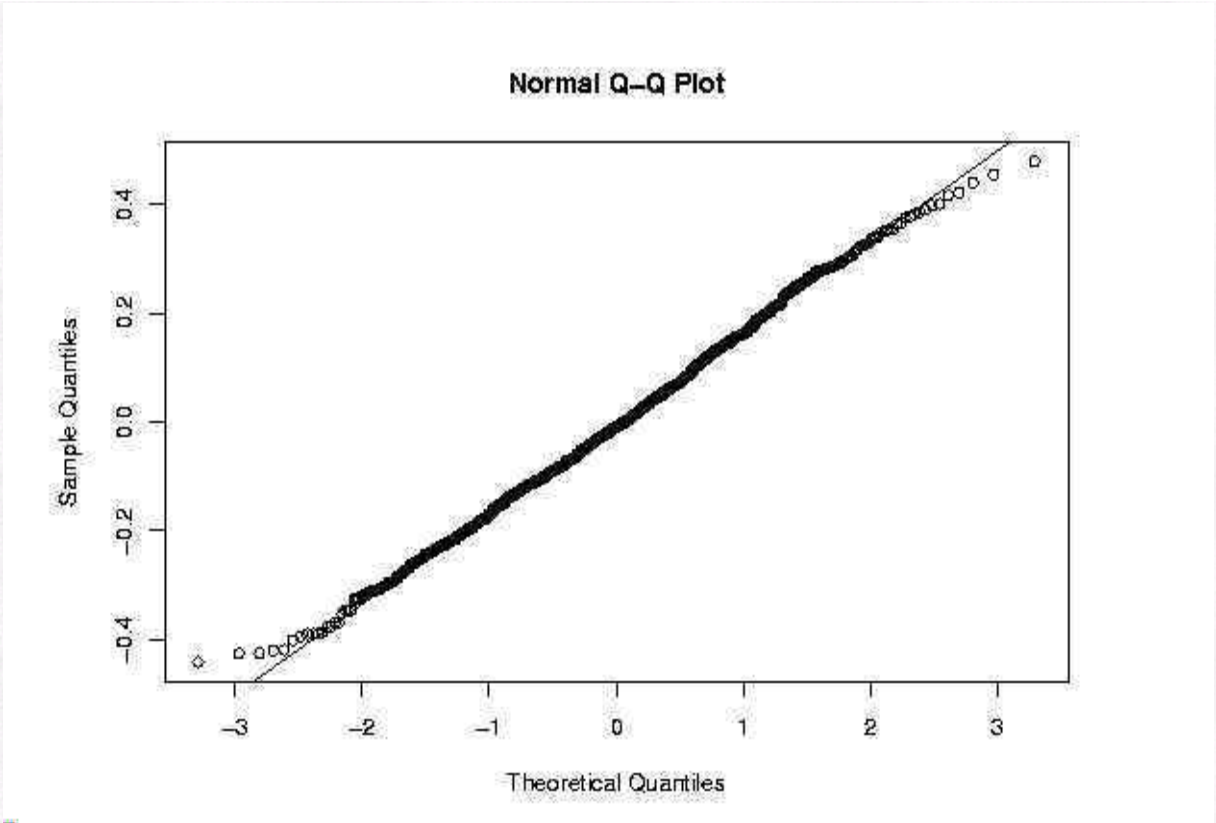
```
> # Results after adding outlier
>
> # Mean and Standard Deviation (standard error)
> # of Bootstrap Samples for Robust Correlation under H0
>
> mean.h0bvec
[1] -0.003922042
> stdev.h0bvec
[1] 0.1657244
>
> # normal theory standard error
> 1/(length(z.x)^.5)
[1] 0.1581139
>
> # Mean and Standard Deviation (standard error)
> # of Bootstrap Samples for Robust Correlation under H1
>
> mean.h1bvec
[1] 0.3206665
> stdev.h1bvec
[1] 0.1531287
> mean.h1bvec-pop.cor
[1] -0.07933351
```

```
> ### Pearson r and Bias
>
> cor.empirical
[1] -0.07661869
> cor.empirical-pop.cor
[1] -0.4766187
>
> ### Robust Correlation and Bias
>
> diff.empirical
[1] 0.3373465
> diff.empirical-mean.hlbvec
[1] 0.01667996
> diff.empirical-pop.cor
[1] -0.06265355

> # Bootstrap empirical power for two-tail test
> # of the robust correlation
>
> power.twotail
[1] 0.494
>
> # Bootstrap empirical p-value for the
> # robust correlation
>
> pvalue.empirical
[1] 0.041

> # Two alternative ways of calculating confidence
> # intervals for Robust Correlation (H0 is preferred)
>
> # Confidence intervals based on the
> # bootstrap sampling distribution for the null
> # and alternate sampling distribution
>
> list(h0.ci = h0.ci)
$h0.ci
[1] 0.02247948 0.66213757

> list(h1.ci=h1.ci)
$h1.ci
[1] 0.009956474 0.614598842
```



References

Hall P., & Wilson S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757-762.

Wilcox, Rand (1997). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, New York.