Correlation Correction of Sample Measures from Bivariate Distributions

Austen Head
Advisor: Jo Hardin
Department of Mathematics
Pomona College

Spring 2008

Abstract

This article focuses on the correlation of data which contain sampling error. The correlations of measures (such as means or quantiles) generated from small samples routinely underestimate the magnitude of the correlations of the parameters of the populations from which those sample measures come. When looking at sample means, there is a well established correction coefficient to account for such a bias. A similar equation to correct for the bias in the correlation of sample quantiles is derived and analyzed in this article. We conclude that using sample maxima is equivalent to, but less versatile than, using sample quantiles in correlating extrema of variables of interest.

Contents

1 Bivariate Distribution Models

A bivariate distribution is simply a distribution across two variables. Bivariate distributions are often used when people collect data on variables that they think are related. For example, college admissions recognize that there is a relationship between high school standardized tests and college grade point averages. Information about how specific variables are related can be valuable in making such decisions.

1.1 Correlating Measures in Bivariate Distributions

Bivariate data can be analyzed in a number of ways. One measure for paired quantitative variables is correlation. The coefficient of correlation is a measure from -1 to 1 of the strength and direction of a linear relationship. For example, if you measure the temperature (accurately) many times in both Fahrenheit and Celsius simultaneously, the coefficient of correlation for the two temperatures will be 1. That is, the temperature in Fahrenheit is perfectly linearly related to the temperature in Celsius. If one variable decreases as the other increases, then the coefficient of correlation will be negative. If the two variables are entirely unrelated or lack a linear component to their relationship (such as a sinusoidal relationship) then the correlation will be zero.

1.2 Attenuation of Measures

Consider the problem of measuring the correlation between average apple size and leaf size on apple trees. From each of 30 apple trees you might take 10 leaves and 10 apples. To find the correlation of average apple and leaf size you would simply take the average leaf size and average apple size from each tree and find the correlation over those 30 points. The sample average apple size and average leaf size are approximations to the true average apple and leaf size on each tree. As you take more samples (i.e., more than 10), by the central limit theorem you will have a better approximate for the true average sizes, but you will still only have an estimate.

Since there is variance within the sample mean the correlation based on the sample will inherently be more spread than the correlation of the true means, which means that the strength of the linear relation will be less strong than it would be if the true means were known. On average, the correlation of the sample means will under-predict the magnitude of the correlation of the true means. In 1904, Spearman [7] discovered this phenomenon which he called attenuation and derived a correction for it when the reliability of measured data is known (see equation 9). Attenuation does not just occur with the correlation of means: any correlation of measures which are subject to sampling error will be attenuated.

2 Correlation Correction for Sample Means

To provide a foundation to better understand later derivations in this paper, this section provides a derivation to the equation that Spearman (1904) [7] discovered that corrects for the attenuation of the correlation of sample means. Throughout the paper, the following scenario will be used to model the theory and demonstrate its capability. We are going to look at the runs speeds of lizards at two different temperatures, X and Y.

Define the following variables as follows: (L for "let")

(L.2.1). $i = 1, \ldots, n_l$ is the i^{th} lizard (n_l is the number of lizards)

(L.2.2). $j = 1, \ldots, n_r$ is the j^{th} run for the i^{th} lizard (let the number of runs n_r be the same for all i)

(L.2.3). x_{ij} is one observation of the random variable X (i.e., the speed at which lizard i runs on its j^{th} run at temperature X).

(L.2.4). $\bar{x_i}$ is a sample mean of the run speeds for the i^{th} lizard at temperature X

(L.2.5). μ_{X_i} is the true mean of the run speeds for the i^{th} lizard at temperature X

(L.2.6). μ_X is the true mean around which μ_{X_i}s are distributed

And similarly for temperature Y. (X and Y will be capitalized for population parameters and in lowercase for sample variables)

There is error $\varepsilon_{x_{ij}}$ in the measurement within each lizard (error estimating μ_{X_i} since the lizards run at different speeds on each run) which will be modeled as:

$$x_{ij} = \mu_{X_i} + \varepsilon_{x_{ij}} \tag{1}$$

The runs x_{ij} are distributed around μ_{X_i}, and the μ_{X_i} are distributed around μ_X. For the duration of this paper, both the distributions of the errors and the distributions of the individual true mean run speeds will be assumed to be Gaussian as:

(A.2.1). $\varepsilon_{x_{ij}} \overset{iid}{\sim} N(0, \sigma_{X_i})$ where $\sigma^2_{X_i}$ is within variance $\quad (\Rightarrow x_{ij} \overset{iid}{\sim} N(\mu_{X_i}, \sigma_{X_i}))$

(A.2.2). $\mu_{X_i} \overset{iid}{\sim} N(\mu_X, \sigma_X)$ where σ^2_X is between variance

(A.2.3). $\varepsilon_{x_{ij}}$ and μ_{X_i} are independent

And define

(L.2.7). $s^2_{x_i}$ is a sample variance within lizard i at temperature X estimated by a dataset

(L.2.8). s^2_x is a sample variance between lizards at temperature X estimated by a dataset

(L.2.9). cor is the correlation of two variables (it is a parameter)

(L.2.10). $\widehat{cor}$ is an estimate of the correlation from a dataset (it is a statistic)

In order to come up with a correction coefficient, we need to define the correlation of the true means in terms of a function of the correlation of the sample means. We want to find a function f such that

$$cor(\mu_{X_i}, \mu_{Y_i}) = f(cor(\bar{x_i}, \bar{y_i})) \tag{2}$$

By the definition of correlation for two random variables A and B

(D.2.1). $cor(A,B) = cov(A,B)/\sqrt{var(A)var(B)}$

We will derive both $cov(\bar{x_i}, \bar{y_i})$ and $var(\bar{x_i})$ by using the model in equation (1).

2.1 Derivation of the Covariance for Sample Means

By definition of covariance, for two random variables A and B

(D.2.2). $cov(A,B) = E[(A - E[A])(B - E[B])]$

We will refer to this definition to solve $cov(\bar{x_i}, \bar{y_i})$. From our model (1) we replace the $\bar{x_i}$ with $\frac{1}{n_r}\sum_{j=1}^{n_r}(\mu_{X_i} + \varepsilon_{x_{ij}})$, and we can simplify to get

$$cov(\bar{x_i}, \bar{y_i}) =$$

$$\frac{1}{n_r^2}\sum_{j=1}^{n_r}\sum_{k=1}^{n_r}(E[\mu_{X_i}\mu_{Y_i} - \mu_{X_i}\varepsilon_{y_{ik}} - \mu_{X_i}\mu_Y - \varepsilon_{x_{ij}}\mu_{Y_i} + \varepsilon_{x_{ij}}\varepsilon_{y_{ik}} + \varepsilon_{x_{ij}}\mu_Y - \mu_X\mu_{Y_i} + \mu_X\varepsilon_{y_{ik}} + \mu_X\mu_Y]) \quad (3)$$

By assumption (A.2.1) (which implies $E[\varepsilon_{x_{ij}}] = E[\varepsilon_{y_{ij}}] = 0$), and (A.2.2) (which implies that $E[\mu_{X_i}] = \mu_X, E[\mu_{Y_i}] = \mu_Y$), and (A.2.3) (independence of errors and true means) we are left with

$$cov(\bar{x_i}, \bar{y_i}) = E[\mu_{X_i}\mu_{Y_i}] - \mu_X\mu_Y = cov(\mu_{X_i}, \mu_{Y_i}) \quad (4)$$

2.2 Derivation of the Variance for Sample Means

To derive the variance of $\bar{x_i}$, we again replace $\bar{x_i}$ with $\frac{1}{n_r}\sum_{j=1}^{n_r}(\mu_{X_i} + \varepsilon_{X_{ij}})$ and by assumptions (A.2.1) and (A.2.2) and (A.2.3) we get

$$var(\bar{x_i}) = var(\mu_{X_i}) + \frac{1}{n_r^2}\sum_{j=1}^{n_r} var(\varepsilon_{X_{ij}}) = \sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r} \quad (5)$$

2.3 The Correction Coefficient for Sample Means

Since we have solved for the covariance (section 2.1) and variances (section 2.2) of the sample means, we can write $cor(\bar{x_i}, \bar{y_i})$ in terms of $cor(\mu_{X_i}, \mu_{Y_i})$ and other parameters using equations (4) and (5):

$$\begin{aligned}
cor(\bar{x_i}, \bar{y_i}) &= \frac{cov(\bar{x_i}, \bar{y_i})}{var(\bar{x_i})var(\bar{y_i})} \\
&= \frac{cov(\mu_{X_i}, \mu_{Y_i})}{\sqrt{\left(\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}\right)\left(\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r}\right)}} \\
&= \frac{cov(\mu_{X_i}, \mu_{Y_i})}{\sqrt{\sigma_X^2\sigma_Y^2}} \frac{\sqrt{\sigma_X^2\sigma_Y^2}}{\sqrt{\left(\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}\right)\left(\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r}\right)}} \\
&= cor(\mu_{X_i}, \mu_{Y_i}) \sqrt{\frac{\sigma_X^2\sigma_Y^2}{\left(\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}\right)\left(\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r}\right)}}
\end{aligned}$$

So we have the following relation:

$$cor(\bar{x}_i, \bar{y}_i) = cor(\mu_{X_i}, \mu_{Y_i}) \sqrt{\left(\frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}}\right)\left(\frac{\sigma_Y^2}{\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r}}\right)} \tag{6}$$

which we can rewrite as

$$cor(\mu_{X_i}, \mu_{Y_i}) = cor(\bar{x}_i, \bar{y}_i) \sqrt{\left(\frac{\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}}{\sigma_X^2}\right)\left(\frac{\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r}}{\sigma_Y^2}\right)} \tag{7}$$

Since we can estimate $cor(\bar{x}_i, \bar{y}_i)$ and all of the variances from the data, we can estimate the correlation of the true means as:

$$\widehat{cor}(\mu_{X_i}, \mu_{Y_i}) = \widehat{cor}(\bar{x}_i, \bar{y}_i) \sqrt{\left(\frac{s_X^2 + \frac{s_{X_i}^2}{n_r}}{s_X^2}\right)\left(\frac{s_Y^2 + \frac{s_{Y_i}^2}{n_r}}{s_Y^2}\right)} \tag{8}$$

Remember though that this equation requires each of the assumptions stated earlier in this section. The equation that Spearman [7] derived in 1904 is equivalent, but is stated as

$$cor(\mu_{X_i}, \mu_{Y_i}) = \frac{cor(\bar{x}_i, \bar{y}_i)}{\sqrt{rel_{X_i} rel_{Y_i}}} \tag{9}$$

where rel_{X_i} is the reliability of the observations of the random variable X. Reliability is defined as the true variance divided by the measured variance (for means, the true variance is the variance between, and the measured variance is the variance of the mean, i.e., variance between plus variance within divided by number of samples). So equation (7) is equivalent to (9) in the case of normality.

2.4 Extrapolations on the Correction Coefficient

Following the same procedure that we did to arrive at equation (6) we can say in general for sample measures a and b that estimate the true parameters A and B that the following holds. (In fact, the following holds for any A, B, a, and b as long as $cor(A, B)$ and $cor(a, b)$ are defined, but this paper does not require such generality.)

$$cor(a, b) = cor(A, B) \left(\frac{cov(a, b)}{cov(A, B)}\right) \sqrt{\left(\frac{var(A)}{var(a)}\right)\left(\frac{var(B)}{var(b)}\right)} \tag{10}$$

or writing $cor(A, B)$ in terms of the estimates:

$$cor(A, B) = cor(a, b) \left(\frac{cov(A, B)}{cov(a, b)}\right) \sqrt{\left(\frac{var(a)}{var(A)}\right)\left(\frac{var(b)}{var(B)}\right)} \tag{11}$$

This relation between the correlation of sample measures and correlation of true measures will be used throughout the paper. If we can estimate the parameters from the observations, and if we can derive the ratio of the covariance of the sample measures to the covariance of the true measures and the ratio of the variance of sample measures to the variance of true measures, then we can estimate the correlation of the true measures.

In section 9 this general relation between the correlation of sample measures and the correlation of true measures will be discussed in the case where the types of measures used in the correlation are different (for example measuring the correlation between max run speed and mean muscle mass).

2.5 Criticisms of Correction

Winne and Belfry (1982) [8] criticize this correction method because it requires true variances which can only be approximated. In underestimates of reliability (i.e., underestimates of the ratio of true variance to measured variance) the correction coefficient will over-correct on average. Winne and Belfry suggest that a confidence band be put around the initial observed correlation and that the correction formula should be applied to the end points of this band (using the method that Forsyth and Feldt [3] put forth). They conclude that although corrected coefficients of correlation can exceed 1 (which is the theoretical limit) people should not be deterred from using the correction as long as the estimates for reliability are sound.

3 Information on Maxima and How Different Variables Affect Their Correlation

Although section 2 focused on the correlation of means, there are many instances where researchers are more interested in extreme performances. For example, when looking at lizard run speeds, it is more interesting to biologists to note how fast they can run when they are running their fastest. Extreme performances represent the speeds that lizards would run at to catch prey or escape predators. This section will use maxima to estimate extreme performance, see section 6 for an analysis of using quantiles to estimate extreme performances.

To discuss maxima we must first give some definitions

(L.3.1). M_{X_i} is the "true" max run speed for the i^{th} lizard

(L.3.2). max_{x_i} is the sample max run speed for the i^{th} lizard over the n_r runs

From the sample max and "true" max we set up our model for maxima as

$$max_{x_i} = M_{X_i} - \epsilon_{x_i} \tag{12}$$

where ϵ_{x_i} is the error resulting when the "true" max is not caught by the sample. And for this paper the following assumptions will be made

(A.3.1). The errors in the max model at the two temperatures are independent of each other (ϵ_{x_i} and ϵ_{y_i} are independent)

(A.3.2). M_{X_i} and ϵ_{x_i} are independent of each other (which implies max_{x_i} and M_{X_i} are dependent on each other for Equation (12) to hold), and M_{Y_i} and ϵ_{y_i} are independent

(A.3.3). M_{X_i} and ϵ_{y_i} are independent, and ϵ_{x_i} and M_{Y_i} are independent

You will notice that we are putting "true" max in quotes; this is because for some distributions there is no true max. For example in a normal distribution $lim_{n_r \to \infty} max_{x_i} = \infty$ and so there is no true max. Despite this, for now we are going to maintain (A.2.1) that the runs are distributed normally, but we will say that the amount of runs in the "population" of runs is finite but sufficiently large to obtain the "true" max of the distribution. Note that for almost all situations that we might be interested in, there are only a finite number of times that we would be interested in (i.e., a lizard can only run a finite number of times in its lifetime).

A derivation relating to maxima can just as appropriately be used for minima with trivial changes.

3.1 Distribution of Maxima

There are three types of Extreme Value Distributions: type 1 are unbounded above and below (Gumbel Distribution), type 2 are unbounded above and bounded below (Fréchet Distribution), type 3 are bounded above and unbounded below (reverse Weibull Distribution) (Coles, 2001 [1]). These three can all be represented by the Generalized Extreme Value Distribution

$$F(z) = exp\left(-\left(1 + \xi\left(\frac{z-a}{b}\right)\right)^{-\frac{1}{\xi}}\right) \tag{13}$$

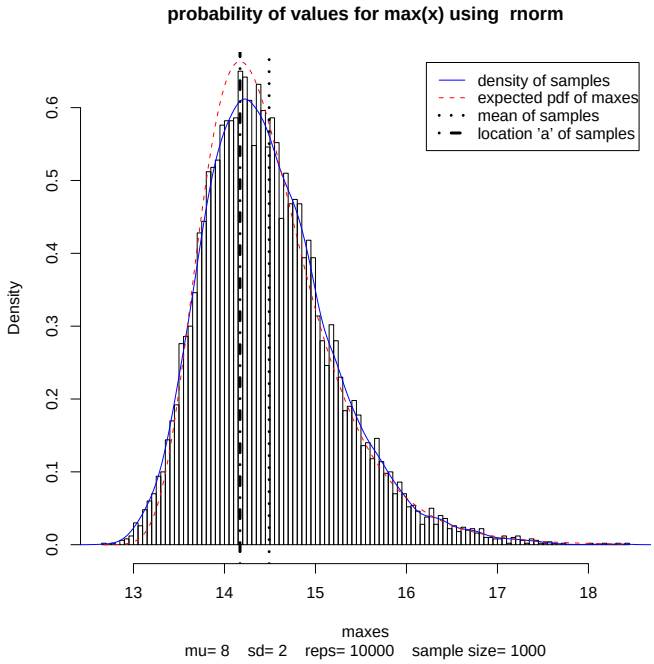

Figure 1: Histogram demonstrates how maxima of sample size 1000 are distributed and approximately follow the $GEV1$ pdf. Notice that the estimated density (solid blue curve) is still not quite the same as a $GEV1$ pdf (dotted red curve).

with associated probability density function

$$f(z) = \frac{1}{b}\left(1+\xi\left(\frac{z-a}{b}\right)\right)^{-\frac{1}{\xi}-1} exp\left(-\left(1+\xi\left(\frac{z-a}{b}\right)\right)^{-\frac{1}{\xi}}\right) \tag{14}$$

where a is the location parameter ($a \in \mathbb{R}$ is the mode), b is the scale parameter ($b > 0, var(z) = b^2\pi^2/6$), ξ is the shape parameter ($\xi \in \mathbb{R}$, $\lim_{\xi\to 0}$ corresponds to type 1, $\xi > 0$ corresponds to type 2, and $\xi < 0$ corresponds to type 3). If the initial distribution is normal (as we are assuming the runs are by (A.2.1)) then the limiting distribution for the maximal extreme is a $GEV1$ distribution (Generalized Extreme Value type 1 distribution) (Kotz, 2000 [6]). See Figure 1 and the histograms on the axes in Figures 2 and 3.

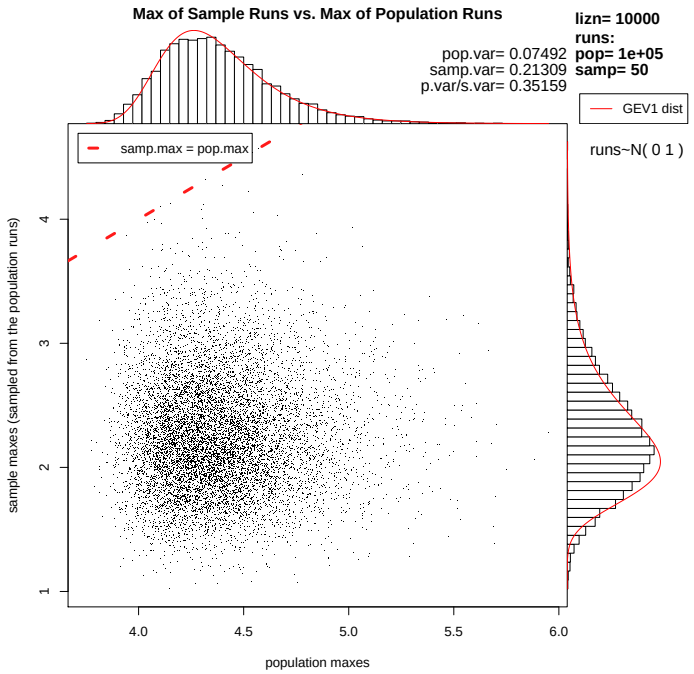

Figure 2: Scatter Plot demonstrates how maxima from samples of size 50 runs per lizard (on the Y axis) of a population of size 1,000,000 runs per lizard and the population maxima (on the X axis) are distributed and approximately follow the $GEV1$ pdf. The distribution of the histogram population on the X axis more accurately follows the $GEV1$ distribution since the sample size of the "population" from which the max is taken is much larger.

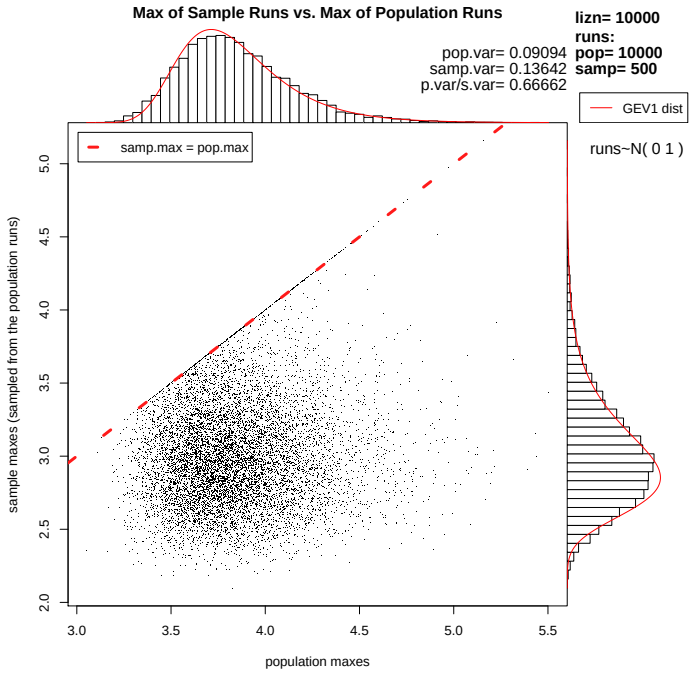

Figure 3: If the population size is not many times larger than the sample size then the true max is often caught by the sample as evidenced by the points which fall on the red dotted line (y=x). Here the sample size is 500 runs from a population of 10,000 runs per lizard.

3.2 Factors affecting the Correlation of Sample Maxima

Using normally distributed data (as described in assumptions (A.2.1), (A.2.2), and (A.2.3)) we have created plots from simulated data to display how different variables affect $cor(max_{x_i}, max_{y_i})$. After running simulations it appears that $cor(max_{x_i}, max_{y_i})$ is directly related to $cor(\mu_{X_i}, \mu_{Y_i})$ (see Figure 4). The accuracy of the linear model demonstrated by Figures 4 and 5 support the hypothesis that $cor(\mu_{X_i}, \mu_{Y_i})$ is very closely related to $cor(max_{x_i}, max_{y_i})$ (at least for normally distributed runs). After seeing that this is the case, we investigated each of the variables mentioned in Figure 4.

We will remark on how each of the variables affects $cor(max_{x_i}, max_{y_i})$ in the captions of the graphs. The character ρ will denote $cor(\mu_{X_i}, \mu_{Y_i})$ and it is investigated in Figures 4 and 5. The number of runs n_r is investigated in Figures 6 and 7. The variance of the runs within the lizards $\sigma^2_{X_i}, \sigma^2_{Y_i}$ and between the lizards σ^2_X, σ^2_Y is investigated in Figure 8.

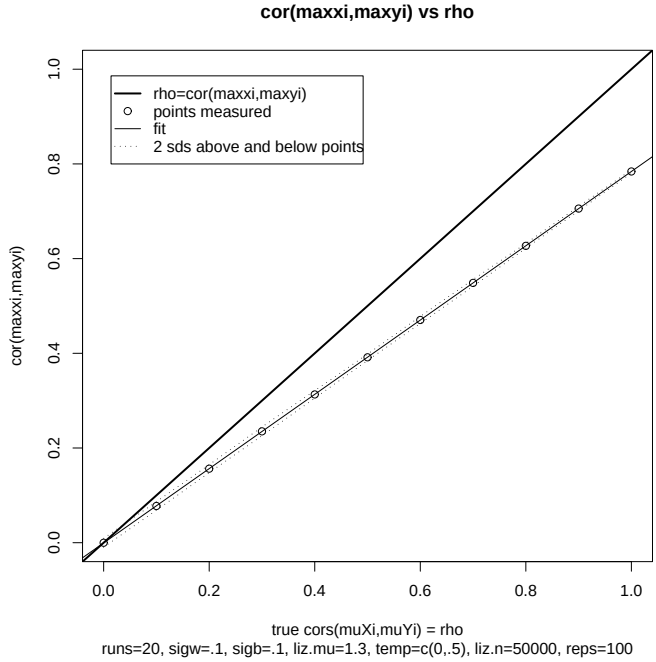

Figure 4: Scatter Plot demonstrating how the $cor(max_{x_i}, max_{y_i})$ (on the Y axis) is related to ρ $(= cor(\mu_{X_i}, \mu_{Y_i}))$. The other variables in the plot are held constant where the number of runs n_r=20, variance within $\sigma^2_{X_i}$=.1, variance between σ^2_X=.1, true means μ_X=1.3 and μ_Y=1.8, and the number of lizards n_l=50,000. This simulation was repeated 100 times to calculate a distribution of the correlation of the sample maxima for each value of the correlation of the true means.

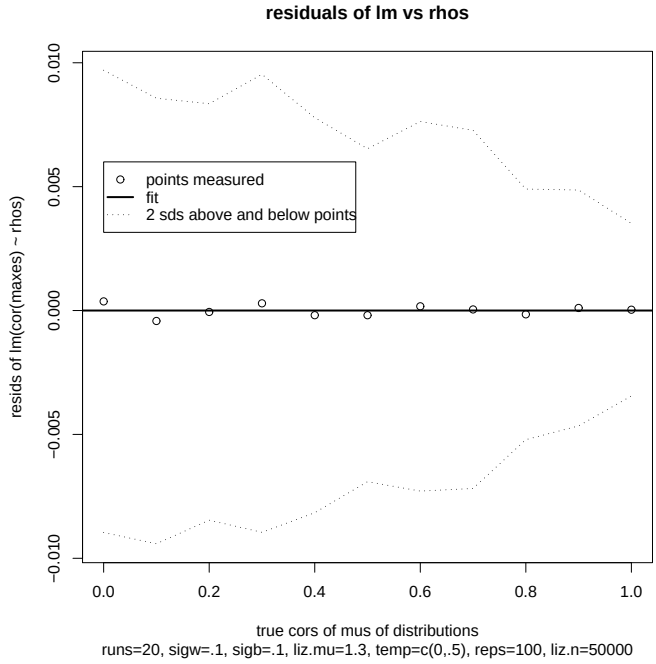

Figure 5: Residuals from the linear model in Figure 4. Note that these seem to be randomly distributed and are very small. It looks as though there is some heteroscedasticity with the variance decreasing as ρ increases, but that $cor(max_{x_i}, max_{y_i})$ is linearly related to ρ.

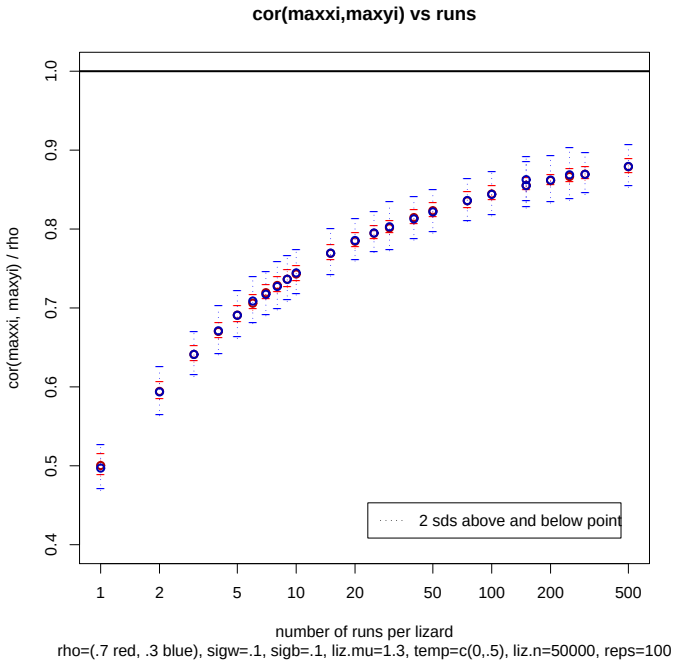

Figure 6: Scatter Plot demonstrates how increasing the number of runs per lizard, n_r, increases the $cor(max_{x_i}, max_{y_i})/\rho$ with diminishing gains for each added run. Note that the scale on the Y axis is a ratio of $cor(max_{x_i}, max_{y_i})/\rho$ so it is possible to compare two different ρ values simultaneously. The difference in the standard deviations of the 100 repetitions of $cor(max_{x_i}, max_{y_i})$s at each point (demonstrated by the dotted lines in the plot) for the different ρs is largely due to the fact that the Y axis is scaled by the division by ρ.

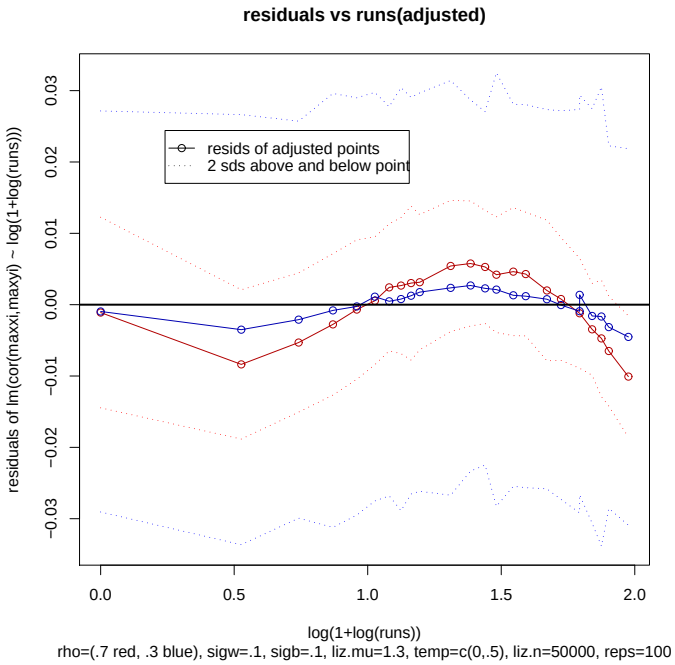

Figure 7: Residuals from an attempt to linearize the model in Figure 6. It looks as though $ln(1 + ln(n_r))$ is a good model in that the residuals are very small, but the residuals show a clear pattern which suggest it is not actually correct. This model was derived by guess and check and is only of heuristic value.

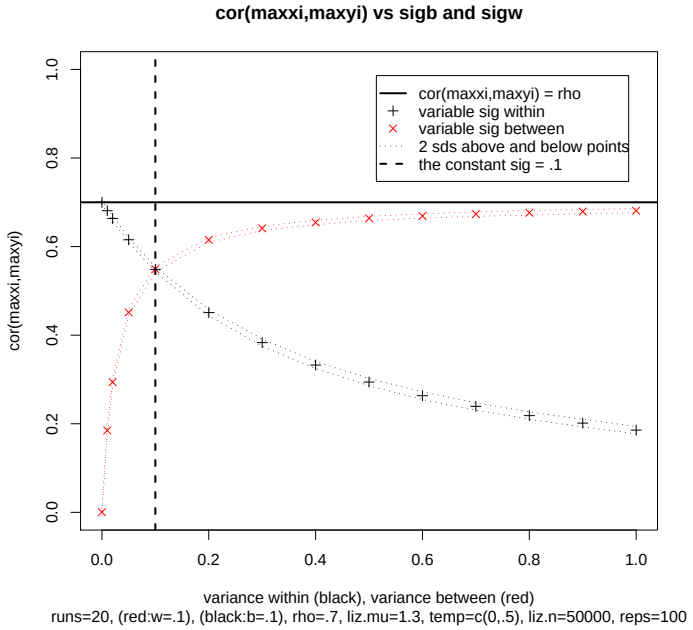

Figure 8: Scatter Plot of how $\sigma^2_{X_i}$ and σ^2_X affect $cor(max_{x_i}, max_{y_i})$. In black the between variance is set at $\sigma^2_X = 0.1$ and the within variance $\sigma^2_{X_i}$ increases from 0 to 1. In red, the within variance is set at $\sigma^2_{X_i} = 0.1$ and the between variance increases from 0 to 1. As $\sigma^2_{X_i}$ (variance within individual lizards) increases $cor(max_{x_i}, max_{y_i})$ decreases. This makes sense because if almost all the error is within variance, then the correlation should be very low since the differences between the lizards seem relatively smaller. Note that when $\sigma^2_{X_i} = 0$ every run speed will be exactly μ_{X_i} and so the sample max will equal the true mean (and the true max) which we see on this graph since the black $+$ at $\sigma^2_{X_i} = 0$ is at the same value as ρ (the horizontal line at 0.7). As σ^2_X increases along the red curve, $\sigma^2_{X_i}$ gets relatively smaller by comparison so $cor(max_{x_i}, max_{y_i})$ increases.

4 Correlation Correction for Sample Maxima

Following the example for deriving the correction to the correlation of sample means from section 2 we will solve the covariance and variances of the sample maxima in terms of the true maxima.

4.1 Derivation of the Covariance for Sample Maxima

Solving for the covariance of sample maxima we can use the model in equation (12) and the definition of covariance (D.2.2) to get

$$\begin{aligned} cov(max_{x_i}, max_{y_i}) &= E[(max_{x_i} - E[max_{x_i}])(max_{y_i} - E[max_{y_i}])] \\ &= E[((M_{X_i} - \epsilon_{x_i}) - E[M_{X_i} - \epsilon_{x_i}])((M_{Y_i} - \epsilon_{y_i}) - E[M_{Y_i} - \epsilon_{y_i}])] \end{aligned} \quad (15)$$

By assumptions for the maxima ((A.3.1), (A.3.2), and (A.3.3)) this simplifies to

$$\begin{aligned} cov(max_{x_i}, max_{y_i}) &= E[M_{X_i} M_{Y_i}] - E[M_{X_i}]E[M_{Y_i}] \\ &= cov(M_{X_i}, M_{Y_i}) \end{aligned} \quad (16)$$

So we find that $cov(max_{x_i}, max_{y_i}) = cov(M_{X_i}, M_{Y_i})$ similar to what we found in equation (4) with the sample means and true means.

4.2 Derivation of the Variance for Sample Maxima

Again replacing max_{x_i} with $M_{X_i} - \epsilon_{x_i}$ and by (A.3.2) we have

$$\begin{aligned} var(max_{x_i}) &= var(M_{X_i} - \epsilon_{x_i}) \\ &= var(M_{X_i}) + var(\epsilon_{x_i}) \end{aligned} \quad (17)$$

But unfortunately that does not help us much since we do not know the distribution or the variance of ϵ_{x_i}.

4.3 The Correction Coefficient for Sample Maxima

After defining the covariance of the true maxima and the variance of the true maxima in terms of the sample maxima we get:

$$\begin{aligned} cor(max_{x_i}, max_{y_i}) &= \frac{cov(max_{x_i}, max_{y_i})}{\sqrt{var(max_{x_i})var(max_{y_i})}} \\ &= \frac{cov(M_{X_i}, M_{Y_i})}{\sqrt{(var(M_{X_i}) + var(\epsilon_{x_i}))(var(M_{Y_i}) + var(\epsilon_{y_i}))}} \\ &= cor(M_{X_i}, M_{Y_i}) \sqrt{\frac{(var(M_{X_i}))(var(M_{Y_i}))}{(var(M_{X_i}) + var(\epsilon_{x_i}))(var(M_{Y_i}) + var(\epsilon_{y_i}))}} \end{aligned} \quad (18)$$

But since we do not know the value of $var(\epsilon_{x_i})$ or $var(M_{X_i})$, equation (18) does not simplify any further and we will instead write it as

$$cor(max_{x_i}, max_{y_i}) = cor(M_{X_i}, M_{Y_i}) \sqrt{\frac{var(M_{X_i})var(M_{Y_i})}{var(max_{x_i})var(max_{y_i})}} \quad (19)$$

Or rearranging the equation,

$$cor(M_{X_i}, M_{Y_i}) = cor(max_{x_i}, max_{y_i}) \sqrt{\frac{var(max_{x_i}) var(max_{y_i})}{var(M_{X_i}) var(M_{Y_i})}} \tag{20}$$

Recall that this equation relies on assumptions about the covariance which only hold when all the assumptions for maxima hold (see section 3). Although we could theoretically use data to estimate the correlation of the true maxima as

$$\widehat{cor}(M_{X_i}, M_{Y_i}) = \widehat{cor}(max_{x_i}, max_{y_i}) \sqrt{\frac{\widehat{var}(max_{x_i}) \widehat{var}(max_{y_i})}{\widehat{var}(M_{X_i}) \widehat{var}(M_{Y_i})}} \tag{21}$$

We do not have any way to estimate the variance of the true maxima, so we do not have $\widehat{var}(M_{X_i})$ or $\widehat{var}(M_{Y_i})$.

5 Simulated Analysis of the Correction Coefficient for Maxima

5.1 From a Large Finite Normal Population with a Known $var(M_{X_i})$

From Equation (21) we have:

$$\widehat{cor}(M_{X_i}, M_{Y_i}) = \widehat{cor}(max_{x_i}, max_{y_i})\sqrt{\frac{\widehat{var}(max_{x_i})\widehat{var}(max_{y_i})}{\widehat{var}(M_{X_i})\widehat{var}(M_{Y_i})}}$$

for any initial distribution (not just normal). In this equation we can calculate the numerator from the sample maxima. In this section, we will demonstrate how effectively this equation estimates $cor(M_{X_i}, M_{Y_i})$ for different numbers of runs if we know the variance of the true maxima.

In order to demonstrate the effectiveness of equation (20) we have simulated a dataset that includes 33 lizards each running 1,000,000 times. The data were created with the assumptions of normality that were used in the section about the means (section 2). Hence, we will use notation from (A.2.1) and (A.2.2) to describe the dataset. The true mean run speeds at the two temperatures are $\mu_X = 1.3, \mu_Y = 1.3 + 0.5 = 1.8$. The between lizard variance is $\sigma^2_X = \sigma^2_Y = 0.7$ (so $\mu_{X_i} \overset{iid}{\sim} N(1.3, 0.7)$). The within variance is $\sigma^2_{X_i} = \sigma^2_{Y_i} = 0.1$ (so $x_{ij} \overset{iid}{\sim} N(\mu_{X_i}, 0.1)$). The true individual means were constructed to have a correlation of 0.7 (i.e., $cor(\mu_{X_i}, \mu_{Y_i}) = 0.7$). The parameters that we used all come from Professor Stephen Adolph in the Harvey Mudd College Biology department who recommended them to us based on data that he has collected by running lizards on a track.

Since we know the values of the variances of the true maxima for this dataset, we can use equation (20) to estimate the correlation of true maxima. By resampling from the simulated dataset we can simulate as many lizards as we want each running up to 1,000,000 times. Because this is a finite simulated population, we are able to find the maximum value of each lizard's run speed to calculate $var(M_{X_i}), var(M_{Y_i}), cor(M_{X_i}, M_{Y_i})$. So we can compare the corrected correlation of the sample maxima (assuming the exact values of the variances of the true maxima is known) to the correlation of the true maxima.

We first resampled 2,000 lizards from the initial 33 lizards and 2 random runs from the 1,000,000 runs for each of the 2,000 lizards. We then found the true max run speeds of those 2,000 lizards so we would know the value of the correlation of the true 2,000 maxima at each temperature. We then compared the correlation of the maxima of the 2 runs for each of the 2,000 lizards at the two different temperatures (i.e., the uncorrected correlation of the sample maxima) to the correlation of the true maxima. The whole process to arrive at the ratio $(cor(max_{x_i}, max_{y_i})/cor(M_{X_i}, M_{Y_i}))$ for 2 runs per lizard for the 2,000 lizards was then repeated 2,000 times to get a distribution of those ratios. The mean of these 2,000 repetitions can be seen on Figure 9 as the bottom left point – the X axis gives the number of runs (2 in this case), the Y axis gives the ratio of the sample correlation over the true correlation, and the color black (the bottom curve) represents using the uncorrected correlation of the sample maxima (i.e., not using equation (20)). This single point demonstrates that on average (with parameters as described above), if you calculate the correlation of the sample maxima of two runs, the correlation will only be about 85% of the correlation of the true maxima of those lizards. This entire process was repeated for up to 1,000 runs and is represented by the black points (the bottom curve) on Figure 9.

Each uncorrected correlation of sample maxima was also corrected using equation (20). Remember this is only possible because we constructed the dataset and so were able to exactly calculate the variance of the true maxima. The corrected correlations of sample maxima were also compared to the correlations of the true maxima, and this is represented on Figure 9 by the red points (the top curve).

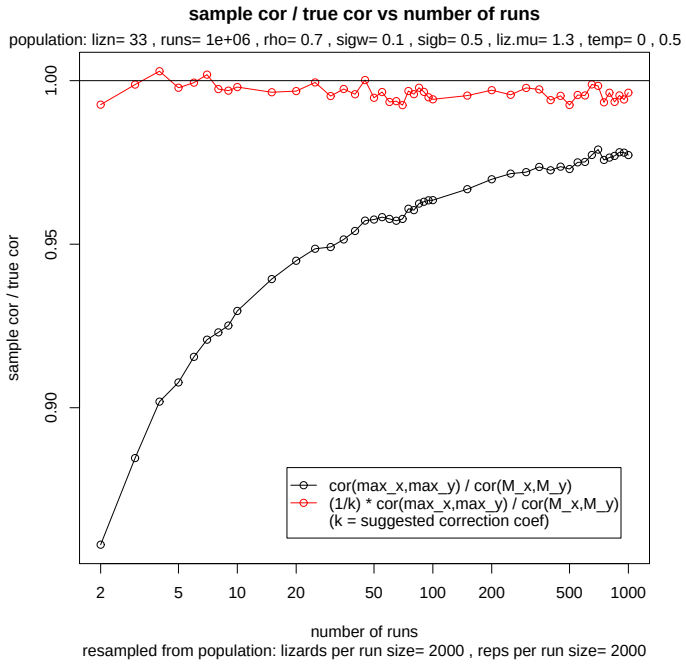

Figure 9: This graph demonstrates that if we are able to estimate the variance of the true maxima, then the correction coefficient with the correlation of sample maxima improves the estimation of the correlation of the true maxima. The curves represents the ratio of the correlation of sample maxima (red is corrected, black is uncorrected) to the correlation of true maxima as a function of the number of runs in the sample. As the number of runs is increased, the correlation of sample maxima more accurately reflects the correlation of true maxima. The corrected correlation of sample maxima fairly accurately predicts the correlation of true maxima for any number of runs.

From Figure 9 we see that for the parameters used in this dataset, on average you need about 30 runs to get 95% of correlation of the true maxima if you are not using the correction coefficient. Using the correction coefficient in equation (20) we get a much better estimate starting at 2 runs. With large samples, the correction does not improve the estimate of the correlation of the true maxima as much as it does with small samples, but from the graph it is evident that even with as many as 1000 samples using the correction coefficient still improves the estimate. In short, if we know or have a good estimate of the variance of the true maxima, and if the assumptions in section 3 hold, then this figure demonstrates that using the correction coefficient will give an improved estimate for the correlation of the true maxima. Refer to section 2.5 to see criticisms of correcting for the correlation of sample means which also apply to the correcting of the correlation of sample maxima.

6 Correlation Correction for Sample Quantiles

6.1 Selecting a Definition for Quantiles

There are several definitions for quantiles: the statistical software R has a quantile function which lets the user select from nine different definitions of quantiles. The following information is taken directly from the R Help for Package Stats (Frohne, 2008) [4]. Types 1-3 are for discrete distributions and 4-9 are for continuous distributions. The default method for R is type 7.

- Type 1: Inverse of empirical distribution function.
- Type 2: Similar to type 1 but with averaging at discontinuities.
- Type 3: SAS definition: nearest even order statistic.
- Type 4: p(k) = k / n. That is, linear interpolation of the empirical cdf.
- Type 5: p(k) = (k - 0.5) / n. That is a piecewise linear function where the knots are the values midway through the steps of the empirical cdf. This is popular amongst hydrologists.
- Type 6: p(k) = k / (n + 1). Thus p(k) = E[F(x[k])]. This is used by Minitab and by SPSS.
- Type 7: p(k) = (k - 1) / (n - 1). In this case, p(k) = mode[F(x[k])]. This is used by S.
- Type 8: p(k) = (k - 1/3) / (n + 1/3). Then p(k) ≈ median[F(x[k])]. The resulting quantile estimates are approximately median-unbiased regardless of the distribution of x.
- Type 9: p(k) = (k - 3/8) / (n + 1/4). The resulting quantile estimates are approximately unbiased for the expected order statistics if x is normally distributed.

Because we will be assuming normality as in section 2, it makes sense to use Type 9 quantiles. In other words we will say that the kth ordered sample out of n total samples represents the quantile

$$p = \frac{k - 3/8}{n + 1/4} \tag{22}$$

However, because we are dealing only with normal distributions, we can estimate quantiles by taking the sample mean and the sample standard deviation. If the ith lizard runs any number of times, we can estimate the pth quantile by

$$p = \bar{x_i} + s_{x_i} * q_p \tag{23}$$

where q_p is the pth quantile of a $N(0, 1)$ distribution.

In section 6.2 we will compare the advantages and disadvantages of quantiles as defined in equations (22) and (23), and then use estimated sample quantiles (as in equation (23)) for the remainder of the paper.

6.2 Reasons to Use Quantiles

Since we are assuming normality, if we look at n runs and pick the max, that is equivalent to finding the $\frac{n-3/8}{n+1/4}$ quantile (see section 6.1 for an explanation of quantiles). But assuming normality, we can also estimate a quantile with the mean of the sample plus the sample standard deviation times the pth quantile of a normal(0,1) distribution.

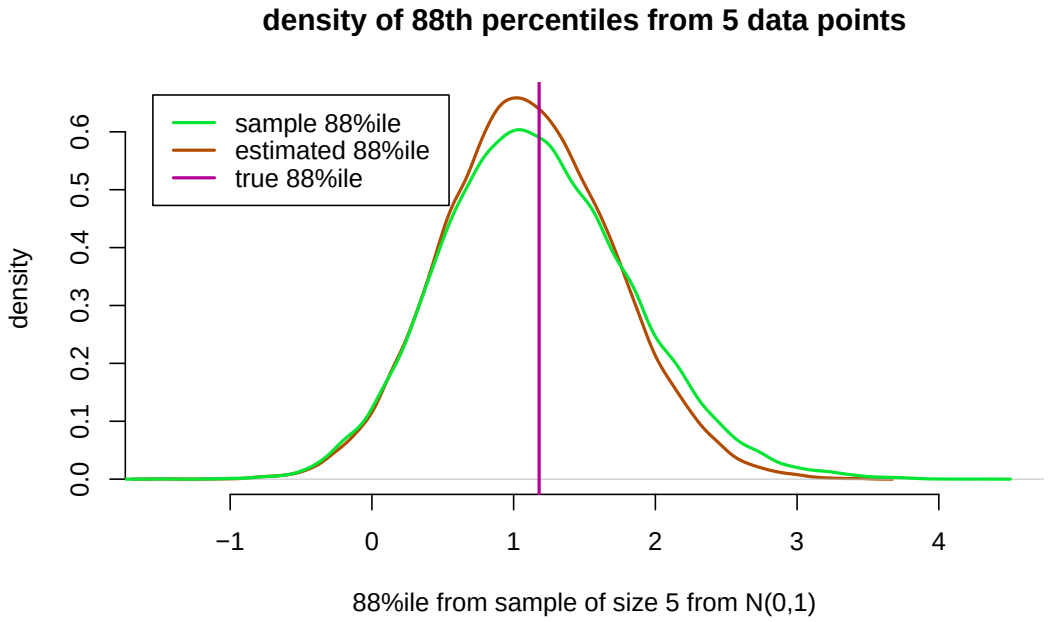

Figure 10: (Although the graph says 88%, the exact value $\frac{5-3/8}{5+1/4}$ was actually used to create the graph). The sample $\frac{5-3/8}{5+1/4} \approx .88$ quantile (green) is the distribution of the maximum sample from 5 samples of a normal(0,1) distribution. The estimated .88 quantile in red is taken from the sample mean of the 5 runs and the sample standard deviation. The density graph above was created by taking 5 samples from a normal $N(0,1)$ distribution 30,000 times. The true .88 quantile of N(0,1) is represented by the purple vertical line.

In Figure 10 we have taken 5 samples from a N(0,1) 30,000 times. The green curve represents the distribution of the sample maxima of the 5 samples which is equivalent to the $\frac{5-3/8}{5+1/4}$ ($\approx .88$) quantile. The red curve represents the .88 quantile estimated from the sample mean and standard deviation of the same 5 runs. Note that the estimated .88 quantile is distributed nearly the same as the sample quantile, and in fact the estimated quantile has less variance. Remember that we are assuming normality when we estimate quantiles with mean and standard deviation.

However, suppose you only have 5 samples and you want to estimate the .95 or the .99 quantile. In Figures 11 and 12 we see that since there are only 5 runs, the max is still estimating the .88 quantile, but if we estimate quantiles with the sample mean and sample standard deviation then we can get an estimate of other quantiles. When researchers are taking the max of some number of samples, that is equivalent to taking the $\frac{n-3/8}{n+1/4}$ quantile of that distribution. Since researchers are often more interested in optimal performance when they are measuring maxima, then it makes much more sense to use quantiles (especially if the initial distribution is normal). Using this method of quantile estimation allows researchers to determine any quantile of performance, whereas looking at maxima only gives the $\frac{n-3/8}{n+1/4}$ quantile.

In essence, taking the sample maxima is a way to estimate a quantile. Keeping this in mind, there are three very important reasons to estimate quantiles for data with small sample sizes. First, we can estimate any quantile with the sample mean and sample standard deviation. This means researchers can be more precise in their level of optimal performance. When researchers only have 5

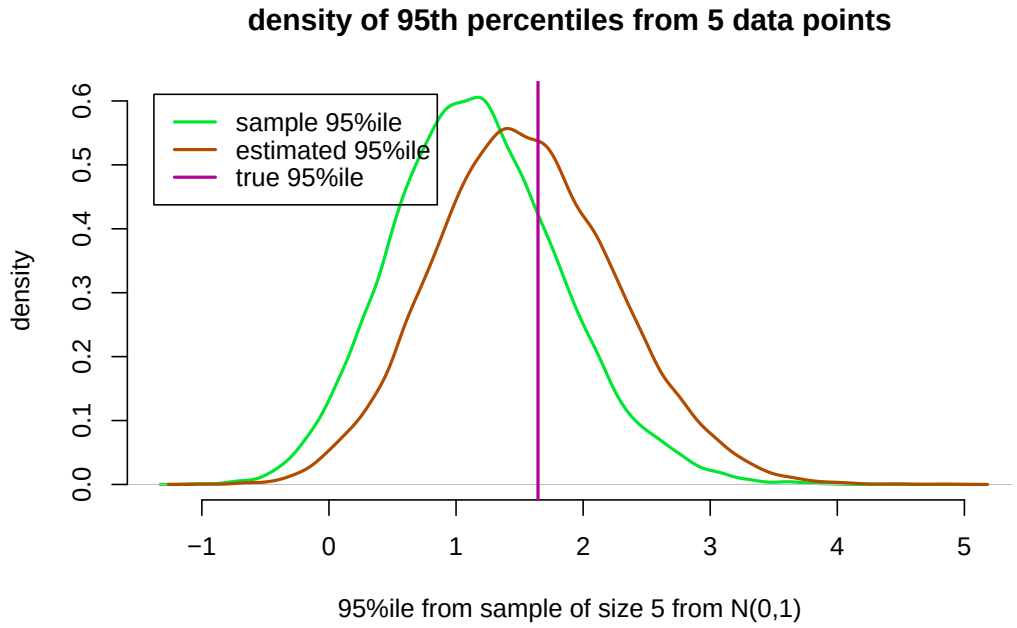

Figure 11: This is a similar graph to Figure 10 except estimating the .95 quantile using the sample mean and sample standard deviation. We see that by using the estimated quantile we can get an estimate of any quantile measure of performance whereas the sample max can estimate the $\frac{n-3/8}{n+1/4}$ quantile. The density graph above was created by taking 5 samples from a normal $N(0,1)$ distribution 30,000 times.

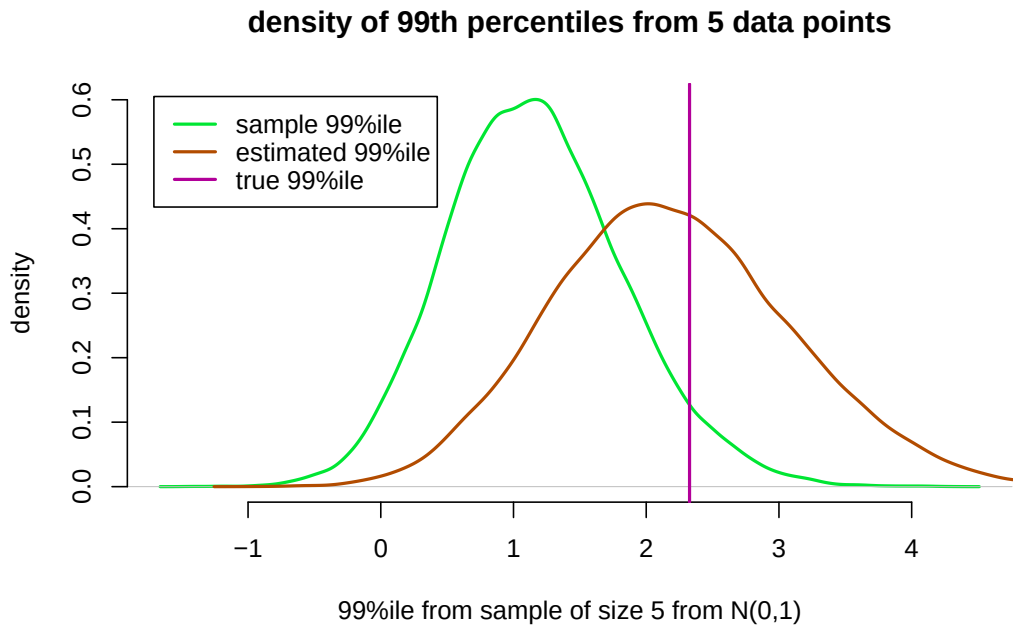

Figure 12: Again, the same graph as above except estimating the .99 quantile. Again we see that the distribution of the estimated quantile is a good estimate for this measure. The density graph above was created by taking 5 samples from a normal $N(0,1)$ distribution 30,000 times.

samples from which to take the max, they are locked in to only taking the .88 quantile, and it takes 12 samples to even reach the .95 quantile. Using the mean and standard deviation and the fact that the data are normally distributed, any sample of size > 1 can be used to estimate a true quantile (although of course larger sample sizes will give more precise estimates of those true quantiles). Second, in section 3, we noted that there is no "true" max for a normal distribution, but there is a finite quantile value for any distribution. Thus estimating quantiles for normally distributed data is something that actually makes sense. In fact, the "true" max is actually just the quantile $\lim_{n\to\infty} \frac{n-3/8}{n+1/4} = 1$ (i.e., the 100% quantile). Third, we can derive a closed form equation to correct for the attenuation in the correlation of sample quantiles (see the rest of this section) where we know or can estimate everything we need for the correction, whereas the correction for the correlation of sample maxima requires the variance of the true maxima which we often do not know and cannot estimate.

6.3 Defining the Quantile More Precisely

Let the variables and assumptions be defined as before (in section 2), and let

(L.6.1). q_p = the p quantile of a normal $N(0,1)$ distribution.

With this we are assuming that the runs are approximately normally distributed, but because q_p is a constant it can be replaced later to fit a more appropriate model if necessary.

Recall the assumption from section 2:

(A.2.1). $\varepsilon_{x_{ij}} \overset{iid}{\sim} N(0, \sigma_{X_i})$ where σ_{X_i} is within variance $\quad (\Rightarrow x_{ij} \overset{iid}{\sim} N(\mu_{X_i}, \sigma_{X_i}))$

(A.2.2). $\mu_{X_i} \overset{iid}{\sim} N(\mu_X, \sigma_X)$ where σ_X is between variance

(A.2.3). $\varepsilon_{x_{ij}}$ and μ_{X_i} are independent

Recall our model from equation (1).

$$x_{ij} = \mu_{X_i} + \varepsilon_{X_{ij}}$$

So the .99 quantile in the distribution of runs for lizard i is approximated by:

$$\bar{x_i} + s_{x_i} * q_{.99} \tag{24}$$

Also recall

(D.2.2). $cor(A,B) = cov(A,B)/\sqrt{var(A)var(B)}$

Using the model in equation 23

$$p = \bar{x_i} + s_{x_i} * q_p$$

we can again solve for the covariance and the variances of the true measures in terms of the sample measures.

6.4 Derivation of the Covariance for Sample Quantiles

By definition of covariance for three random variables A, B, and C where A and B are independent (D.6.2). $cov(A+B,C) = cov(A,C) + cov(B,C)$

Since $\bar{x}$ and s_{x_i} are independent we can apply (D.6.2):

$$\begin{aligned} & cov(\bar{x_i} + s_{x_i} * q_p, \bar{y_i} + s_{y_i} * q_p) \\ &= cov(\bar{x_i}, \bar{y_i}) + cov(\bar{x_i}, s_{y_i} * q_p) + cov(s_{x_i} * q_p, \bar{y_i}) + cov(s_{x_i} * q_p, s_{y_i} * q_p) \\ &= cov(\bar{x_i}, \bar{y_i}) \qquad (25) \\ &= cov(\mu_{X_i}, \mu_{Y_i}) \qquad (26) \end{aligned}$$

See equation (4) for explanation how equation (26) is derived from equation (25).

6.5 Derivation of the Variance for Sample Quantiles Assuming Normality

Under normality the distribution of the sample variances is chi squared. Note that the result is not asymptotic based on sample size.

$$\begin{aligned} (n_r-1)s_{x_i}^2/\sigma_{X_i}^2 \sim \chi_{n_r-1}^2 \quad &\Rightarrow \quad var((n_r-1)s_{x_i}^2/\sigma_{X_i}^2) = 2(n_r-1) \\ &\Rightarrow \quad \frac{(n_r-1)^2}{\sigma_{X_i}^4} var(s_{x_i}^2) = 2(n_r-1) \\ &\Rightarrow \quad var(s_{x_i}^2) = \frac{2\sigma_{X_i}^4}{n_r-1} \qquad (27) \end{aligned}$$

The Delta Method [5] states: Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. for a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}(g(Y_n) - g(\theta)) \to N(0, var(Y_n\sqrt{n})(g'(\theta))^2) \qquad (28)$$

in distribution.

In the Delta Method let $Y_n = s_{x_i}^2$, $\theta = \sigma_{X_i}^2$, $g(b) = \sqrt{b}$, $g'(b) = \frac{1}{2}(b)^{-1/2}$. By the central limit theorem we know $s_{x_i}^2 \to$ normal ($s_{x_i}^2$ is asymptotically normal as you increase the number of runs), so it is safe to use the Delta Method here, and we have

$$\begin{aligned} \sqrt{n}(g(Y_n) - g(\theta)) \quad &\to \quad N(0, var(Y_n\sqrt{n})(g'(\theta))^2) \\ \sqrt{n_r}(s_{x_i} - \sigma_{X_i}) \quad &\to \quad N(0, var(s_{x_i}\sqrt{n_r})(\frac{1}{2}(\sigma_{X_i}^2)^{-1/2})^2) \\ &\to \quad N(0, var(s_{x_i}^2)\frac{n_r}{4\sigma_{X_i}^2}) \\ &\to \quad N(0, (\frac{2\sigma_{X_i}^4}{n_r-1})(\frac{n_r}{4\sigma_{X_i}^2})) \quad \text{see equation (27)} \\ &\to \quad N(0, \frac{n_r\sigma_{X_i}^2}{2(n_r-1)}) \qquad (29) \end{aligned}$$

$$\Rightarrow \quad var(\sqrt{n_r}(s_{x_i} - \sigma_{X_i}) = \frac{n_r \sigma_{X_i}^2}{2(n_r - 1)}$$

$$\Rightarrow \quad var(s_{x_i}) = \frac{\sigma_{X_i}^2}{2(n_r - 1)} \tag{30}$$

And from equation (5) we have

$$var(\bar{x_i}) = var(\mu_{X_i}) + \frac{1}{n_r^2} \sum_{j=1}^{n_r} var(\varepsilon_{X_{ij}}) = \sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}$$

From equations (5) and (30) we have:

$$var(\bar{x_i} + s_{x_i} * q_p) = var(\bar{x_i}) + q_p^2 var(s_{x_i}) + q_p cov(\bar{x_i}, s_{x_i}) \tag{31}$$

$$= \sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r} + \frac{q_p^2 \sigma_{X_i}^2}{2(n_r - 1)} + q_p * 0 \tag{32}$$

$$= \sigma_X^2 + \frac{\sigma_{X_i}^2 (n_r(2 + q_p^2) - 2)}{2n_r(n_r - 1)} \tag{33}$$

Because the variance of the quantile is derived using the Delta method, we need to be careful when using small sample sizes.

David 1970 [2] gives another derivation of the variance of the sample standard deviation from a normal distribution which is exact and does not require the delta method, but which is more complicated.

$$var(s_{x_i}) = \frac{\sigma_{X_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r - 1}{2})} \right) \tag{34}$$

Where Γ is the standard gamma function.

Applying David's derivation of the variance of the sample standard deviation we can get the exact value for the variance of the sample quantile:

$$var(\bar{x_i} + s_{x_i} * q_p) = \sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r} + q_p^2 \frac{\sigma_{X_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r - 1}{2})} \right) \tag{35}$$

6.6 The Correlation for Sample Quantiles Assuming Normality

As in the section on the correction of the correlation of sample means and the correction of the correlation of sample maxima, we can use the equations of the variance and covariance of the sample quantile to produce the following correction equation (keeping in mind the assumptions of normality from section 6.3)

$$cor(\bar{x_i} + s_{x_i} q_p, \bar{y_i} + s_{y_i} q_p) =$$

$$\frac{cor(\mu_{X_i} + \sigma_{X_i} q_p, \mu_{Y_i} + \sigma_{Y_i} q_p) \sigma_{X_i} \sigma_{Y_i}}{\sqrt{\left(\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r} + q_p^2 \frac{\sigma_{X_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r - 1}{2})} \right) \right) \left(\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r} + q_p^2 \frac{\sigma_{Y_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r - 1}{2})} \right) \right)}} \tag{36}$$

Let

$$\begin{aligned} Vq_x &= s_x^2 + \frac{s_{x_i}^2}{n_r} + q_p^2 \frac{s_{x_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r-1}{2})} \right) \\ Vq_y &= s_y^2 + \frac{s_{y_i}^2}{n_r} + q_p^2 \frac{s_{y_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r-1}{2})} \right) \end{aligned}$$

Then we can estimate

$$\widehat{cor}(\mu_{X_i} + \sigma_{X_i} q_p, \mu_{Y_i} + \sigma_{Y_i} q_p) = \widehat{cor}(\bar{x}_i + s_{x_i} q_p, \bar{y}_i + s_{y_i} q_p) \frac{\sqrt{Vq_x Vq_y}}{s_{x_i} s_{y_i}} \tag{37}$$

7 Simulated Analysis of the Correction Coefficient for Quantiles

7.1 From a Large Finite Normal Population

From equation (37) we have:

$$\widehat{cor}(\mu_{X_i} + \sigma_{X_i} q_p, \mu_{Y_i} + \sigma_{Y_i} q_p) = \widehat{cor}(\bar{x}_i + s_{x_i} q_p, \bar{y}_i + s_{y_i} q_p) \frac{\sqrt{V q_x V q_y}}{s_{x_i} s_{y_i}}$$

We are able to estimate all of the values necessary to estimate the correlation of the true sample pth quantile.

Using the same dataset described in section 5.1 and the same process to create the graph in Figure 9, we have created a similar graph comparing the uncorrected and corrected correlation of sample .99 quantiles. In Figure 13 each point is a ratio of the correlation of sample .99 quantiles over the correlation of the true .99 quantile (which we are able to find since the data is simulated).

Figure 13 demonstrates that if we follow assumptions from section 2 and estimate the variance of the true quantiles with the variance of the sample quantiles, then using the correction coefficient in Equation (37) gives an estimate of the correlation of the true quantile. In Figure 13 we chose to look at the .99 quantile as opposed to the .95, .75, or .50 because the .99 quantile as estimated by the sample mean and sample standard deviation is more variable (in Figures 11 and 12, compare the variability of the red curves which are the quantiles estimated from the mean and standard deviation) and so a quantile closer to the median would produce even better estimates. The improvement in estimating the correlation of the true quantiles that equation (37) provides (especially for a small number of runs) can be seen in the curves on Figure 13.

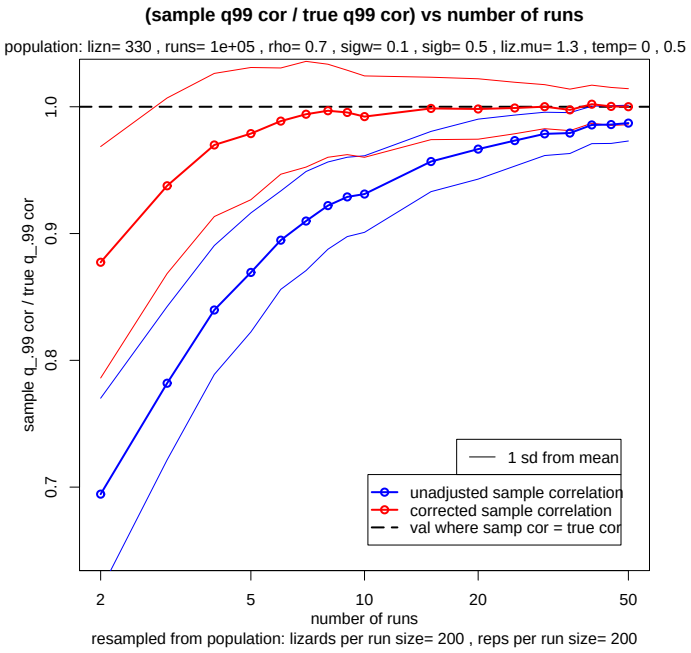

Figure 13: This graph demonstrates that from the data alone, we can use the correction coefficient in equation (37) to get a less biased estimate of the correlation of the true quantiles. The curves represents the ratio of the correlation of sample .99 quantile (corrected and uncorrected) to the correlation of true .99 quantiles as a function of the number of runs in the sample. As the number of runs is increased, the correlation of sample quantiles more accurately reflects the correlation of true quantiles. The corrected correlation of sample quantiles is a much better predictor of the correlation of true quantiles for any number of runs.

8 Comparing the Use of Quantiles and Maxima

Given the equations for the correction of the correlation of sample maxima (21):

$$\widehat{cor}(M_{X_i}, M_{Y_i}) = \widehat{cor}(max_{x_i}, max_{y_i})\sqrt{\frac{\widehat{var}(max_{x_i})\widehat{var}(max_{y_i})}{\widehat{var}(M_{X_i})\widehat{var}(M_{Y_i})}}$$

and the correction of the correlation of sample quantiles (37):

$$\widehat{cor}(\mu_{X_i} + \sigma_{X_i}q_p, \mu_{Y_i} + \sigma_{Y_i}q_p) = \widehat{cor}(\bar{x}_i + s_{x_i}q_p, \bar{y}_i + s_{y_i}q_p)\frac{\sqrt{Vq_xVq_y}}{s_{x_i}s_{y_i}}$$

we can immediately see an advantage to the latter since we are able to estimate all the necessary parts of that equation. Additionally, by considering the advantages of quantiles as outlined in section 6.2 we can see the added versatility of quantiles and their inherent usefulness in defining optimal performance that sample maxima do not provide. The only case in which maxima should be used in preference over quantiles is if the true maximum values or the variance of the true maxima (as needed in equation (21)) are already known or are estimable.

9 Correlations of Two Different Measures

As we saw in section 2.4 equation (11):

$$cor(A,B) = cor(a,b)\left(\frac{cov(A,B)}{cov(a,b)}\right)\sqrt{\left(\frac{var(a)}{var(A)}\right)\left(\frac{var(b)}{var(B)}\right)}$$

We took advantage of this to estimate the correlation of true measures with sample measures for means, quantiles, and maxima. But, this equation can be used even more widely. In fact, we can let A and B be different types of measures. For example, continuing with the lizards example, we can let A be the mean run speed at temperature X (i.e., $\mu_{X_i} = A$) and B be the lower 15th quantile of run speed at temperature Y (i.e., $\mu_{Y_i} + \sigma_{Y_i} q_{.15} = B$). We can then let a and b be the sample mean and sample quantile which estimate A and B. As it turns out, using the same techniques as in earlier sections, $cov(a,b) = cov(A,B)$ for A and B as any of the mean, max, or quantile and a and b the sample measures of A and B. So in our example:

$$\widehat{cor}(\mu_{X_i}, \mu_{Y_i} + q_{.15}\sigma_{Y_i}) = \widehat{cor}(\bar{x}_i, \bar{y}_i + q_{.15} s_{y_i})\sqrt{\left(\frac{s_x^2 + s_{x_i}^2/n_r}{s_x^2}\right)\left(\frac{V q_y}{s_y^2}\right)} \tag{38}$$

(Note that by our model for quantiles in equation (23), if we are looking at the mean, that is equivalent to looking at the median since $q_{.5} = 0$.)

This kind of correction between two different measures is particularly important when researchers are interested in different conditions. For example, a biologist might want to find the correlation of the slowest 15% of lizards in relation to their mean daily food consumption. With the knowledge of how correlations of sample measures are related to measures of true measures, researchers can better estimate the correlation of true measures from their samples. As with all correlation correction, we need to be mindful of the criticisms raised in section 2.5.

10 Conclusion

When researchers measure the correlation of sample maxima, they are necessarily measuring the correlation of sample quantiles (see section 8). In normal distributions, sample quantiles can be estimated with the mean and the standard deviation to allow the researcher to look at any quantile rather than being locked into an empirical quantile (see sections 6.1 and 6.2). We have derived a correction to more accurately estimate the correlation of true quantiles (see section 6.6). Since maxima and quantiles can both be used as measures of extreme value, we recommend that instead of measuring the correlation of sample maxima, researchers should use the corrected correlation of sample quantiles (see equation 37) which is:

$$\widehat{cor}(\mu_{X_i} + \sigma_{X_i} q_p, \mu_{Y_i} + \sigma_{Y_i} q_p) = cor(\bar{x}_i + s_{x_i} q_p, \bar{y}_i + s_{y_i} q_p) \frac{\sqrt{Vq_x Vq_y}}{s_{x_i} s_{y_i}}$$

where

$$\begin{aligned} Vq_x &= s_x^2 + \frac{s_{x_i}^2}{n_r} + q_p^2 \frac{s_{x_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r-1}{2})} \right) \\ Vq_y &= s_y^2 + \frac{s_{y_i}^2}{n_r} + q_p^2 \frac{s_{y_i}^2}{n_r} \left(n_r - 1 - \frac{2\Gamma^2(\frac{n_r}{2})}{\Gamma^2(\frac{n_r-1}{2})} \right) \end{aligned}$$

This equation holds when the assumptions in section 2 are valid. The notation can be found in sections 2 and 6.3. More generally, from equation (11) in section 2.4, we have the correlation correction equation of

$$cor(A, B) = cor(a, b) \left(\frac{cov(A, B)}{cov(a, b)} \right) \sqrt{\left(\frac{var(a)}{var(A)} \right) \left(\frac{var(b)}{var(B)} \right)}$$

in which a and b can be used as the sample measures of true measures A and B. These equations correct for the attenuation inherent in the correlation of measures taken with sampling error and give a better estimate of a correlation of the true measures.

11 Applications to Other Fields

Though the previous work was motivated by a problem in biology, measurement error is ubiquitous. To show the broad applications of these methods, examples from other fields will be described. For each field a situation is given in which the conclusions drawn in section 10 could provide useful information to the researchers. For each experiment, we describe the variables that correspond to the equations in section 10. We have not included information about randomization and some other experimental controls.

11.1 Biology

(The example used in throughout this paper.) Fast run speeds are of interest because they represent the speeds lizards will run to escape predators and catch prey. Suppose that there is an expected climate change and you want to find the correlation of fast run speeds of lizards at two different temperatures. The correlation of the .95 quantile of lizard speeds at the two temperatures answers the question: "Is there a linear relationship between how fast lizards can run at one temperature and how fast they can run at another temperature?"

Experiment: Have 15 lizards each run on a track 3 times at each of the two temperatures.
X, Y the two temperatures
x_{ij} the speed that lizard i (i from 1 to n_l) runs on trial j (j from 1 to n_r) at temperature X
n_r the number of runs per lizard at each temperature = 3
n_l the number of lizards = 15

Use the data to estimate the correlation of true .95 quantile run speeds at the two different temperatures.

11.2 Ecology

Ecologists might be interested in the correlation between average monthly rainfall between two large forests.

Experiment: Collect data on rainfall for 12 months at 5 locations in each of two forests.
X, Y the two forests
x_{ij} the rainfall on month i (i from 1 to n_l) at measuring spot j (j from 1 to n_r) at forest X
n_r the number of locations in each forest where you measure rainfall =5
n_l the number of months =12

Use the data to estimate the correlation of the true mean of monthly rainfall over the two forests.

11.3 Economics

An economist may be interested in the average number of weekly commercials and the sales of the product that is being advertised.

Experiment: Show commercials in 15 cities and measure the number of commercials shown and number of sales of the product each week for 5 weeks.
X, Y number of commercials and number of sales
x_{ij} the number of commercials shown in city i (i from 1 to n_l) on week j (j from 1 to n_r).
n_r the number of weeks =5
n_l the number of cities =15

Use the data to estimate the correlation of the true mean number of weekly commercials versus the mean number of sales per week.

11.4 Education

Suppose a school has an advanced track which students must qualify for by performing well in both language and mathematical tasks. Although it might be easier for the school to have one set of advanced students, perhaps there are some gifted students who excel in one discipline but not the other. If there is a significant correlation between performance on language and mathematics at the school then this track might be appropriate, otherwise they might want to consider having separate advanced tracks for language and math.

Experiment: Have 15 students each take 3 different language tests and 3 different math tests.
X, Y language ability and mathematical ability
x_{ij} the performance of student i (i from 1 to n_l) on test j (j from 1 to n_r) on the language test
n_r the number of tests per discipline = 3
n_l the number of students = 15

Use the data to estimate the correlation of the true mean test scores for the two different disciplines.

11.5 Engineering

Suppose there is a machine that requires several of the same component A, and the machine will suffer damage if a component A is damaged too much from regular use. If the machine is used under both strenuous and easy conditions, the engineers might be interested in the relationship between the wear of a component A when the machine is used under strenuous and easy conditions.

Experiment: Run 15 machines under strenuous conditions and then measure the wear on each of 5 A components, replace those components and run the machines under easy conditions and measure the wear on each of the A components.
X, Y strenuous and easy conditions
x_{ij} the wear of the jth (j from 1 to n_r) A component of machine i (i from 1 to n_l) under strenuous conditions
n_r the number of components per machine =5
n_l the number of machines = 15

Use the data to estimate the correlation of true .95 quantile of wear at between the two conditions.

11.6 Psychology

Psychologists might be interested in the relationship between happiness and reported job satisfaction. They might conjecture that reported average weekly job satisfaction is more related to lows in happiness rather than average happiness.

Experiment: Have 15 participants rate both their level of happiness and their level of job satisfaction once per day for a week.
X, Y daily reported happiness and daily job satisfaction
x_{ij} the reported happiness level of participant i (i from 1 to n_l) on day j (j from 1 to n_r)
n_r the number of days per week = 7
n_l the number of participants = 15

Use the data to estimate the correlation of the true .10 quantile of weekly happiness versus the true mean weekly job satisfaction level.

11.7 Sociology

Sociologists might be interested to know if there is some relationship between poverty rate and murder rate in cities. Although this would be simple enough with one year's worth of data (i.e., one value for the poverty rate and one value for the murder rate for each city), we can get a better approximation of the correlation of the mean poverty and murder rates if we use data from several years.

Experiment: For 10 cities, find the murder rate and the poverty rate for the past 5 years
X, Y murder and poverty rates
x_{ij} the number of murders in city i (i from 1 to n_l) in year j (j from 1 to n_r).
n_r the number of years =5
n_l the number of cities =10

Use the data to estimate the correlation of the true mean poverty rate versus the true mean murder rate in cities.

References

[1] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[2] Herbert Aron David. *Order Statistics*. John Wiley & Sons, Inc., 1970.

[3] Robert A. Forsyth and Leonard S. Feldt. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *Educational and Psychological Measurement*, 29(1):61–71, Spring 1969.

[4] Ivan Frohne and Rob J Hyndman. Sample quantiles. In *R Project*. ISBN 3-900051-07-0, 2008.

[5] Robert V. Hogg, Joseph W. McKean, and Allen T Craig. *Introduction to Mathematical Statistics*. Pearson Education, Inc., 6th edition, 2005.

[6] Samuel Kotz and Saralees Nadarajah. *Extreme Value Distributions*. Imperial College Press, 2000.

[7] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, January 1904.

[8] Philip H. Winne and M. Joan Belfry. Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, 19(2):125–134, Summer 1982.