NTIA TM-26-581

U.S. Department of Commerce
National Telecommunications and Information Administration

Institute for Telecommunication Sciences
Technical Memorandum

Analysis of Multiple Noisy Measurements and Their Resamplings via Expected Values of Correlation and Mean-Squared-Error

Jaden Pieper and Stephen D. Voran

March 2026
Boulder,
Colorado
its.ntia.gov

Institute for Telecommunication Sciences
Technical Memorandum
NTIA TM-26-581
March 2026

Analysis of Multiple Noisy Measurements and Their Resamplings via Expected Values of Correlation and Mean-Squared-Error

Jaden Pieper and Stephen D. Voran

United States Department of Commerce

National Telecommunications and Information Administration

Institute for Telecommunication Sciences
325 Broadway, Boulder, CO 80305
its.ntia.gov

This publication is available free of charge from:
https://doi.org/10.70220/hqqs0vzx

About the Institute for Telecommunication Sciences

The Institute for Telecommunication Sciences (ITS) is the research and engineering laboratory of the National Telecommunications and Information Administration (NTIA), an agency of the U.S. Department of Commerce (DoC). ITS manages the telecommunications technology research programs of NTIA and the DoC-owned Table Mountain Radio Quiet Zone. ITS works closely with other NTIA Line Offices to support Administration and Agency needs.

The mission of ITS is to ADVANCE innovation in communications technologies, INFORM spectrum and communications policy for the benefit of all stakeholders, and INVESTIGATE our Nation's most pressing telecommunications challenges through research that employees are proud to deliver.

ITS publishes fundamental science communications as part of its Technology Transfer efforts. We are committed to ensuring that our publications are substantive, technically sound, accurate, and clear. A rigorous peer review process, documented in the " ITS Publications Handbook Volume I: Policies (Third Edition)," safeguards the scientific integrity of ITS technical publications, journal articles, and conference papers. The principles for technical peer review of manuscripts guide parallel peer review processes for the publication of software and datasets.

Principal NTIA Formal Publication Series Published by ITS

NTIA Technical Report (TR): Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

NTIA Technical Memorandum (TM): Technical information typically of less breadth than an NTIA Technical Report. The series includes data, preliminary project results, and information for a specific, limited audience.

NTIA Special Publication (SP): Conference proceedings, bibliographies, course and instructional materials, or major scientific studies mandated by Congress.

NTIA Monograph (MG): A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA Handbook (HB): Information pertaining to technical procedures; reference and data guides; and formal user's manuals that are expected to be pertinent for a long time.

For more information about NTIA fundamental research publications, contact the ITS Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. 303-497-3572 or email ITSinfo@ntia.gov.

Disclaimer

Identification of commercial products or services in this report does not imply recommendation or endorsement by the National Telecommunications and Information Administration, Department of Commerce, and U.S. Government.

Contents

Figures

Analysis of Multiple Noisy Measurements and Their Resamplings via Expected Values of Correlation and Mean-Squared-Error

Jaden Pieper[1] and Stephen D. Voran[1]

Abstract: This memorandum provides mathematical relationships for agreement between true measurand values, means of multiple imperfect measurements, and resampled measurement means. Agreement is quantified here in terms of Pearson correlation coefficient (PCC) and mean-squared error (MSE). The derivations of the key relationships range from simple to rather intricate, but the resulting equations are all simple, intuitive, and easy to use. More specifically, given a number of measurands, a number of measurements per measurand, and some basic variance information, the equations provide the expected PCC (or MSE) between measurand values, means of measurements, and means of resampled measurements. Depending on the measurement context and the specific goals for PCC or MSE, these results can guide readers in deciding on an appropriate number of measurements to gather.

Keywords: bootstrapping, correlation, mean opinion score, mean-squared error, measurement error, MOS, resampling, subjective testing

1. Introduction

This work addresses some details in the general problem of making inherently imperfect, or noisy, measurements of some physical property or quantity, known as the measurand. For this work we assume that the value of the measurand is itself a random process, so that the value of the measurand is not fixed. We consider the case where we make measurements of the measurand for multiple conditions or multiple independent draws from the random process. For each draw from the random process, we make multiple measurements in an attempt to ascertain the true value of the measurand.

Our motivating example is the problem of subjective multimedia quality experiments, where a media file (audio, video, or both) has a given quality, and we seek to measure that quality by playing the file for each subject and collecting a vote from each subject. Some examples of these experiments can be found in [1], [2], [3]. The subjects' votes are noisy measurements of a true underlying quality value that is associated with each file. Averaging more votes results in

[1] The authors are with the U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunication Sciences, Boulder, CO 80305.

better approximations of quality, but increasing the number of votes is expensive in terms of time and resources, so in an experiment there is a limit on how many votes each file receives.

One could also consider a situation where the value of the measurand is time-dependent. In that case, multiple measurement devices may be used to simultaneously measure the measurand. Each measurement device produces a noisy estimate of the true value, so more devices would allow for better approximations of the measurand. However, again, there is a cost associated with measurement devices, so there is a limit to how many can be used. For the purpose of this work, we consider this situation to be equivalent to that of subjective experiments.

For either case, we assume that given a value of the measurand, the measurements are accurate (meaning they have the correct expected value) but they are not precise (they have non-zero variance). This implies that, despite the inherent variance of the measurements, if we collect enough measurements, their average will converge to the true value of the measurand.

In the case of subjective testing, the mean of a set of votes is often called a mean opinion score (MOS). The question arises: "What is the relationship between the MOS values and the true quality values, and how does that relationship change if we gather more votes?" Due to some unique characteristics of subjective testing and MOS values, Pearson correlation coefficient (PCC) is often the most appropriate way to compare MOS results. Mean-squared error (MSE) can also be used, but with some limitations.

Thus, we seek to characterize the relationships between means of measurements and true measurand values using PCC and MSE. Given that the true values are unknowable, this is a somewhat theoretical question. We set up the most general possible problem and develop new mathematical results that capture these relationships. In spite of the theoretical nature of the situation, these new results satisfy our intuition for the situation — the measurement variance is key, the distribution of true values is a factor, and additional measurements improve the agreement with the true values.

Given that access to the true values does not exist in practice, resampling or bootstrapping is a popular, practical, data-driven way to answer questions about the correctness, or, at least, the stability, of the gathered data. This practice allows us to generate multiple means from a set of noisy measurements or, in the case of subjective testing, multiple MOS values from a set of votes. This then begs new analogous questions: "What are the MSE and PCC between original measurement means and the means derived via resampling?" We present new mathematical answers for these questions, and the answers also agree with intuition.

This work was motivated by the relatively recent availability of large speech quality datasets, intended for machine learning (ML) applications. These datasets contain speech files and associated MOS values. Between datasets, and even within datasets, the number of votes used to calculate each MOS value can range from 1 to around 50. These MOS values are the "truth data" or "targets" for ML, but depending on the number of votes involved, the connection to "truth" can range from very loose to quite tight. Combining such data in a harmonious way requires understanding these connections. Our work to formally quantify these connections motivates the mathematical results that follow.

2. Problem Statement

Three indices are prevalent throughout this work. The index i refers to the ith value of the measurand, and i runs from 1 to n_f. In the multimedia subjective testing field, this would be the ith media file, but other examples would be the ith location where an RF power or an air temperature must be measured, or the ith time at which a measurement must be made. Correlations and MSE's are calculated across this index, so they capture the variation in the measurand, along with the noise in the measurements.

In this problem n_m measurements are made for each measurand. The index j denotes these individual measurements, and j runs from 1 to n_m. In the multimedia subjective testing field, j denotes the individual subjects' votes for a single media file. In other cases, j could denote the outputs produced by multiple sensors at a single location and/or time. Because test subjects and sensors are imperfect, the individual subject votes or sensor outputs will show some variation across j.

The index k is the resampling index, and it also runs from 1 to n_m. We resample the measurements (indexed by j) and k denotes position in this new list of n_m measurements. For each value of k, we draw, with equal probability, one of the n_m exising measurement values (or samples). This is sampling with replacement. For example, if $n_m = 4$ and the original measurements were $\{X_{i,j}\} = \{1.1,\ 1.2,\ 0.9,\ 1.3\}$ one resampling of those measurements is $\{Z_{i,j}\} = \{1.2,\ 0.9,\ 1.1,\ 0.9\}$. Here $X_{i,4} = 1.3$ and $Z_{i,4} = 0.9$.

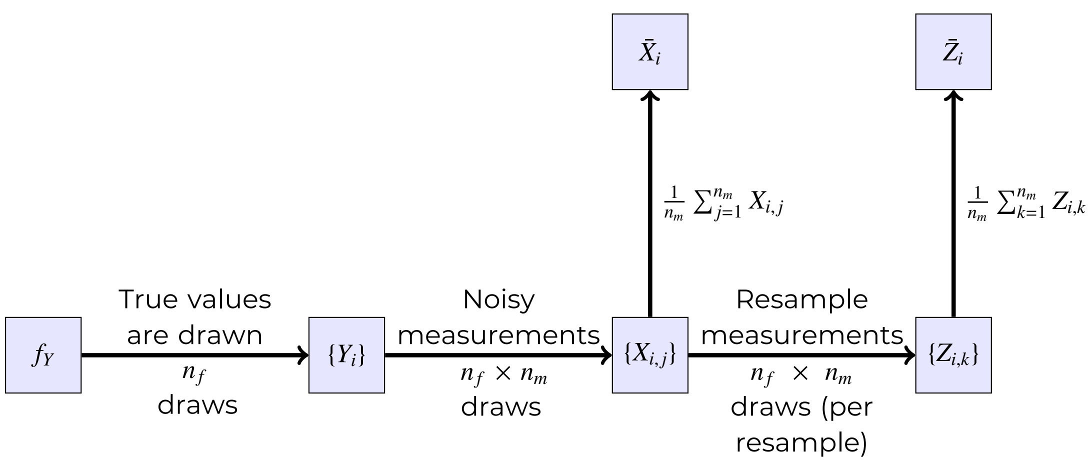

Figure 2.1. Context for the problem showing the chain of events and hence the flow of conditioning.

The context required for the problem statement is shown in Fig. 2.1. This figure depicts the chain of events that we are modeling, and this chain informs the conditioning statements seen in the math that follows.

True values[2] of the quantities to be measured are drawn from an arbitrary, continuous probability distribution f_Y. Let $\{Y_i\}$ be independent and identically distributed (i.i.d.) samples from this distribution.

For each Y_i we have a set of n_m noisy measurements: $X_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,n_m}]$. These are i.i.d. samples, and it is assumed that they respond to the true value, Y_i. In other words we assume

$$\mathbb{E}\left(X_{i,j}|Y_i\right) = Y_i. \tag{1}$$

We also define the conditional variance of the measurements given a value of the measurand as

$$\mathrm{Var}\left(X_{i,j}|Y_i\right) = v_m(Y_i), \tag{2}$$

for all $j = 1, ..., n_m$. Note that the measurement variance is a function of the true value Y_i in order to model real situations. Sensors can have different noise or uncertainty specifications at different parts of the measurement range. In subjective testing, the response variance is highest in the middle of the scale and lower at the extremes of the scale. Examples can be seen in [4]. That is, subjects tend to have higher agreement about stimuli that are very bad or very good, but less agreement about intermediate stimuli.

Resampling is a key step in bootstrapping [5], a technique for investigating the stability of results, often by building confidence intervals. We use resampling here, but we diverge from common practice by asking questions about properties of the bootstrapped statistics in terms of correlation and MSE rather than building confidence intervals. The noisy measurements in the vector X_i are resampled with replacement to produce the vector $Z_i = [Z_{i,1}, Z_{i,2}, \ldots, Z_{i,n_m}]$. Resampling with replacement means that

$$P(Z_{i,k} = X_{i,j}|X_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,n}], Y_i) = \frac{1}{n_m}, \quad \forall j, k = 1, \ldots, n_m. \tag{3}$$

Next we define the sample mean and resampled mean respectively as

$$\bar{X}_i = \frac{1}{n_m} \sum_{j=1}^{n_m} X_{i,j} \ , \tag{4}$$

$$\bar{Z}_i = \frac{1}{n_m} \sum_{k=1}^{n_m} Z_{i,k} \ . \tag{5}$$

We are seeking the population correlation between the true values Y_i and the sample means $\bar{X}_i$, and also the expected sample MSE between these two. In addition, we are seeking the population correlation between the sample mean $\bar{X}_i$ and the resampled mean $\bar{Z}_i$, and also the expected sample MSE between these two. The derivations of these four quantities follow.

[2] These are not available in practice but that does not inhibit the derivations that follow. As a thought experiment, one could consider that “truth” might be *approached* at some great expense. In the multimedia subjective testing case, truth might be approached by an extremely large set of multi-laboratory tests that call upon a large number of subjects and include a large number of media files to establish a stable context. In other cases, truth might be approached with newly calibrated laboratory grade sensors operated with extreme care under ideal conditions.

3. Agreement between Sample Mean and True Value

3.1 Correlation

The population PCC between the sample mean $\bar{X}_i$ and the truth Y_i is defined as

$$\rho(\bar{X}_i, Y_i) = \frac{\mathbb{E}\left(\bar{X}_i Y_i\right) - \mathbb{E}\left(\bar{X}_i\right)\mathbb{E}\left(Y_i\right)}{\sqrt{\text{Var}\left(\bar{X}_i\right)\text{Var}\left(Y_i\right)}}. \tag{6}$$

The expected value of the sample mean can be found through the law of total expectation as follows:

$$\begin{aligned}
\mathbb{E}\left(\bar{X}_i\right) &= \mathbb{E}\left(\frac{1}{n_m}\sum_{j=1}^{n_m} X_{i,j}\right) \\
&= \frac{1}{n_m}\sum_{j=1}^{n_m}\mathbb{E}\left(X_{i,j}\right) \\
&= \frac{1}{n_m}\sum_{j=1}^{n_m}\mathbb{E}\left(\mathbb{E}\left(X_{i,j}|Y_i\right)\right) \\
&= \frac{1}{n_m}\sum_{j=1}^{n_m}\mathbb{E}\left(Y_i\right) \\
&= \mathbb{E}\left(Y_i\right).
\end{aligned} \tag{7}$$

The variance of the sample mean can be found via the law of total variance as follows:

$$\begin{aligned}
\text{Var}\left(\bar{X}_i\right) &= \text{Var}\left(\mathbb{E}\left(\bar{X}_i|Y_i\right)\right) + \mathbb{E}\left(\text{Var}\left(\bar{X}_i|Y_i\right)\right) \\
&= \text{Var}\left(Y_i\right) + \mathbb{E}\left(\text{Var}\left(\bar{X}_i|Y_i\right)\right).
\end{aligned} \tag{8}$$

The second term includes $\text{Var}\left(\bar{X}_i|Y_i\right)$ which can be reduced as follows:

$$\begin{aligned}
\text{Var}\left(\bar{X}_i|Y_i\right) &= \text{Var}\left(\frac{1}{n_m}\sum_{j=1}^{n_m} X_{i,j}|Y_i\right) \\
&= \frac{1}{n_m^2}\sum_{j=1}^{n_m}\text{Var}\left(X_{i,j}|Y_i\right) \\
&= \frac{1}{n_m^2}\sum_{j=1}^{n_m} v_m(Y_i) \\
&= \frac{v_m(Y_i)}{n_m}.
\end{aligned} \tag{9}$$

Using (9) in (8) gives

$$\begin{aligned}
\text{Var}\left(\bar{X}_i\right) &= \text{Var}\left(Y_i\right) + \mathbb{E}\left(\text{Var}\left(\bar{X}_i|Y_i\right)\right) \\
&= \text{Var}\left(Y_i\right) + \mathbb{E}\left(\frac{v_m(Y_i)}{n_m}\right) \\
&= \text{Var}\left(Y_i\right) + \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m}.
\end{aligned} \tag{10}$$

Next we find the expected value of the product of the sample mean and the truth:

$$\begin{aligned}\mathbb{E}\left(\bar{X}_i Y_i\right) &= \mathbb{E}\left(\frac{1}{n_m}\sum_{j=1}^{n_m} X_{i,j} Y_i\right) \\ &= \frac{1}{n_m}\sum_{j=1}^{n_m}\mathbb{E}\left(X_{i,j}Y_i\right) \\ &= \frac{1}{n_m}\sum_{j=1}^{n_m}\mathbb{E}\left(\mathbb{E}\left(X_{i,j}Y_i|Y_i\right)\right) \\ &= \frac{1}{n_m}\sum_{j=1}^{n_m}\mathbb{E}\left(Y_i^2\right) \\ &= \mathbb{E}\left(Y_i^2\right) \\ &= \mathrm{Var}\,(Y_i) + \mathbb{E}\,(Y_i)^2\,, \end{aligned} \tag{11}$$

where we have used the law of total expectation and the definition of variance.

We can now use results (7), (10), and (11) in result (6) to see that the correlation is:

$$\begin{aligned}\rho(\bar{X}_i, Y_i) &= \frac{\mathbb{E}\left(\bar{X}_i Y_i\right) - \mathbb{E}\left(\bar{X}_i\right)\mathbb{E}\,(Y_i)}{\sqrt{\mathrm{Var}\left(\bar{X}_i\right)\mathrm{Var}\,(Y_i)}} \\ &= \frac{\mathrm{Var}\,(Y_i) + \mathbb{E}\,(Y_i)^2 - \mathbb{E}\,(Y_i)^2}{\sqrt{(\mathrm{Var}\,(Y_i) + \frac{1}{n_m}\mathbb{E}\,(v_m(Y_i)))\mathrm{Var}\,(Y_i)}} \\ &= \sqrt{\frac{\mathrm{Var}\,(Y_i)}{\mathrm{Var}\,(Y_i) + \frac{1}{n_m}\mathbb{E}\,(v_m(Y_i))}}. \end{aligned} \tag{12}$$

3.2 Mean-Squared Error

The squared-error between the sample mean $\bar{X}_i$ and the true value Y_i is

$$\epsilon_i^2 = (\bar{X}_i - Y_i)^2. \tag{13}$$

The mean-squared error (MSE), or D^2, is simply the mean of these squared differences,

$$D^2 = \frac{1}{n_f}\sum_{i=1}^{n_f}\epsilon_i^2. \tag{14}$$

This is a random variable and we are seeking its expected value $\mathbb{E}\left(D^2\right)$

$$\begin{aligned}\mathbb{E}\left(D^2\right) &= \mathbb{E}\left(\frac{1}{n_f}\sum_{i=1}^{n_f}\epsilon_i^2\right) = \frac{1}{n_f}\sum_{i=1}^{n_f}\mathbb{E}\left(\epsilon_i^2\right) \\ &= \mathbb{E}\left(\epsilon_i^2\right), \end{aligned} \tag{15}$$

where the final equality exploits the fact the ϵ_i^2 are i.i.d. It follows that

$$\begin{aligned}
\mathbb{E}\left(\epsilon_i^2\right) &= \mathbb{E}\left((\bar{X}_i - Y_i)^2\right) \\
&= \mathbb{E}\left(\mathbb{E}\left((\bar{X}_i - Y_i)^2\right)|Y_i\right) \\
&= \mathbb{E}\left(\mathbb{E}\left((\bar{X}_i - \mathbb{E}\left(\bar{X}_i\right))^2\right)|Y_i\right) \\
&= \mathbb{E}\left(\mathrm{Var}\left(\bar{X}_i|Y_i\right)\right) \\
&= \mathbb{E}\left(\frac{v_m(Y_i)}{n_m}\right) \\
&= \frac{1}{n_m}\mathbb{E}\left(v_m(Y_i)\right),
\end{aligned} \tag{16}$$

where we have used the result (9).

4. Agreement between Sample Mean and Resampled Mean

4.1 Correlation

The population correlation between the sample mean $\bar{X}_i$ and the resampled mean $\bar{Z}_i$ is defined as

$$\rho(\bar{X}_i, \bar{Z}_i) = \frac{\mathbb{E}\left(\bar{X}_i\bar{Z}_i\right) - \mathbb{E}\left(\bar{X}_i\right)\mathbb{E}\left(\bar{Z}_i\right)}{\text{Var}\left(\bar{X}_i\right)\text{Var}\left(\bar{Z}_i\right)}. \tag{17}$$

The expected value of the sample mean $\mathbb{E}\left(\bar{X}_i\right)$ is given in (7), and the expected value of the resampled mean $\mathbb{E}\left(\bar{Z}_i\right)$ is

$$\begin{aligned}
\mathbb{E}\left(\bar{Z}_i\right) &= \mathbb{E}\left(\frac{1}{n_m}\sum_{k=1}^{n_m} Z_{i,k}\right) \\
&= \frac{1}{n_m}\sum_{k=1}^{n_m}\mathbb{E}\left(Z_{i,k}\right) \\
&= \mathbb{E}\left(Z_{i,k}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(Z_{i,k}|X_i, Y_i\right)\right) \\
&= \mathbb{E}\left(\bar{X}_i\right) \\
&= \mathbb{E}\left(Y_i\right).
\end{aligned} \tag{18}$$

The variance of the sample mean $\text{Var}\left(\bar{X}_i\right)$ is given in (8). We use the law of total variance to find the variance of the resampled mean:

$$\begin{aligned}
\text{Var}\left(\bar{Z}_i\right) &= \mathbb{E}\left(\text{Var}\left(\bar{Z}_i|X_i, Y_i\right)\right) + \text{Var}\left(\mathbb{E}\left(\bar{Z}_i|X_i, Y_i\right)\right) \\
&= \mathbb{E}\left(\text{Var}\left(\bar{Z}_i|X_i, Y_i\right)\right) + \text{Var}\left(\bar{X}_i\right) \\
&= \mathbb{E}\left(\text{Var}\left(\bar{Z}_i|X_i, Y_i\right)\right) + \text{Var}\left(Y_i\right) + \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m}.
\end{aligned} \tag{19}$$

The first term includes $\text{Var}\left(\bar{Z}_i|X_i, Y_i\right)$ which can be rewritten as

$$\begin{aligned}
\text{Var}\left(\bar{Z}_i|X_i, Y_i\right) &= \mathbb{E}\left(\left(\bar{Z}_i - \mathbb{E}\left(\bar{Z}_i\right)\right)^2|X_i, Y_i\right) \\
&= \mathbb{E}\left(\bar{Z}_i^2|X_i, Y_i\right) - \mathbb{E}\left(\bar{Z}_i|X_i, Y_i\right)^2 \\
&= \mathbb{E}\left(\bar{Z}_i^2|X_i, Y_i\right) - \bar{X}_i^2.
\end{aligned} \tag{20}$$

The second moment of $\bar{Z}_i$ is

$$\begin{aligned}
\mathbb{E}\left(\bar{Z}_i^2|X_i, Y_i\right) &= \mathbb{E}\left(\frac{1}{n_m^2}\sum_{j=1}^{n_m}\sum_{k=1}^{n_m} Z_{i,j}Z_{i,k}|X_i, Y_i\right) \\
&= \frac{1}{n_m^2}\sum_{j=1}^{n_m}\sum_{k=1}^{n_m}\mathbb{E}\left(Z_{i,j}Z_{i,k}|X_i, Y_i\right) \\
&= \frac{1}{n_m^2}\left(n_m(n_m-1)\mathbb{E}\left(Z_{i,j}|X_i, Y_i\right)^2 + n_m\mathbb{E}\left(Z_{i,j}^2|X_i, Y_i\right)\right) \\
&= \frac{1}{n_m^2}\left(n_m(n_m-1)\bar{X}_i^2 + n_m\mathbb{E}\left(Z_{i,j}^2|X_i, Y_i\right)\right). \qquad (21)
\end{aligned}$$

Next we find find the conditional second moment of a resampled measurement $Z_{i,j}$,

$$\begin{aligned}
\mathbb{E}\left(Z_{i,j}^2|X_i, Y_i\right) &= \sum_{k=1}^{n_m} P(Z_{i,j} = X_{i,k})X_{i,k}^2 \\
&= \frac{1}{n_m}\sum_{k=1}^{n_m} X_{i,k}^2, \qquad (22)
\end{aligned}$$

then use the result (22) in (21) to show

$$\begin{aligned}
\mathbb{E}\left(\bar{Z}_i^2|X_i, Y_i\right) &= \frac{1}{n_m^2}\left(n_m(n_m-1)\bar{X}_i^2 + n_m\frac{1}{n_m}\sum_{k=1}^{n_m} X_{i,k}^2\right) \\
&= \frac{1}{n_m^2}\left(n_m(n_m-1)\bar{X}_i^2 + \sum_{k=1}^{n_m} X_{i,k}^2\right). \qquad (23)
\end{aligned}$$

Using (23) in (20) then gives

$$\begin{aligned}
\text{Var}\left(\bar{Z}_i|X_i, Y_i\right) &= \mathbb{E}\left(\bar{Z}_i^2|X_i, Y_i\right) - \bar{X}_i^2 \\
&= \frac{1}{n_m^2}\left(n_m(n_m-1)\bar{X}_i^2 + \sum_{k=1}^{n_m} X_{i,k}^2\right) - \bar{X}_i^2 \\
&= \frac{1}{n_m^2}\left(\left(\sum_{k=1}^{n_m} X_{i,k}^2\right) - n_m\bar{X}_i^2\right). \qquad (24)
\end{aligned}$$

In order to use this result in (19) we must first take the expected value:

$$\begin{aligned}
\mathbb{E}\left(\text{Var}\left(\bar{Z}_i|X_i, Y_i\right)\right) &= \mathbb{E}\left(\frac{1}{n_m^2}\left(\left(\sum_{k=1}^{n_m} X_{i,k}^2\right) - n_m\bar{X}_i^2\right)\right) \\
&= \frac{1}{n_m^2}\left(\left(\sum_{k=1}^{n_m}\mathbb{E}\left(X_{i,k}^2\right)\right) - n_m\mathbb{E}\left(\bar{X}_i^2\right)\right). \qquad (25)
\end{aligned}$$

We expand the second moment of $\bar{X}_i$ using the definition of variance and the result in (10):

$$\begin{aligned}
\mathbb{E}\left(\bar{X}_i^2\right) &= \text{Var}\left(\bar{X}_i\right) + \mathbb{E}\left(\bar{X}_i\right)^2 \\
&= \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m} + \text{Var}\left(Y_i\right) + \mathbb{E}\left(Y_i\right)^2. \qquad (26)
\end{aligned}$$

Looking back to (25), we see that we also need the second moment of $X_{i,k}$:

$$\begin{aligned}
\mathbb{E}\left(X_{i,k}^2\right) &= \mathbb{E}\left(\mathbb{E}\left(X_{i,k}^2|Y_i\right)\right) \\
&= \mathbb{E}\left(\text{Var}\left(X_{i,k}|Y_i\right) + \mathbb{E}\left(X_{i,k}|Y_i\right)^2\right) \\
&= \mathbb{E}\left(v_m(Y_i)\right) + \mathbb{E}\left(Y_i^2\right) \\
&= \mathbb{E}\left(v_m(Y_i)\right) + \text{Var}\left(Y_i\right) + \mathbb{E}\left(Y_i\right)^2 .
\end{aligned} \tag{27}$$

Using results (26) and (27) in (25) gives

$$\begin{aligned}
\mathbb{E}\left(\text{Var}\left(\bar{Z}_i|X_i, Y_i\right)\right) &= \frac{1}{n_m^2}\left(\left(\sum_{k=1}^{n_m}\mathbb{E}\left(X_{i,k}^2\right)\right) - n_m\mathbb{E}\left(\bar{X}_i^2\right)\right) \\
&= \frac{1}{n_m^2}\left(\left(\sum_{k=1}^{n_m}\mathbb{E}\left(v_m(Y_i)\right) + \text{Var}\left(Y_i\right) + \mathbb{E}\left(Y_i\right)^2\right) - n_m\left(\frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m} + \text{Var}\left(Y_i\right) + \mathbb{E}\left(Y_i\right)^2\right)\right) \\
&= \frac{1}{n_m}\left(\mathbb{E}\left(v_m(Y_i)\right) + \text{Var}\left(Y_i\right) + \mathbb{E}\left(Y_i\right)^2 - \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m} - \text{Var}\left(Y_i\right) - \mathbb{E}\left(Y_i\right)^2\right) \\
&= \left(\frac{n_m - 1}{n_m^2}\right)\mathbb{E}\left(v_m(Y_i)\right) .
\end{aligned} \tag{28}$$

Finally, we can use (28) in (19) to find

$$\begin{aligned}
\text{Var}\left(\bar{Z}_i\right) &= \mathbb{E}\left(\text{Var}\left(\bar{Z}_i|\bar{X}_i, Y_i\right)\right) + \text{Var}\left(Y_i\right) + \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m} \\
&= \left(\frac{n_m - 1}{n_m^2}\right)\mathbb{E}\left(v_m(Y_i)\right) + \text{Var}\left(Y_i\right) + \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m} \\
&= \left(\frac{2n_m - 1}{n_m^2}\right)\mathbb{E}\left(v_m(Y_i)\right) + \text{Var}\left(Y_i\right) .
\end{aligned} \tag{29}$$

We now have four of the five components of the correlation given in (17). It remains to find the expected value of the product of sample mean and resampled mean:

$$\begin{aligned}
\mathbb{E}\left(\bar{X}_i\bar{Z}_i\right) &= \mathbb{E}\left(\mathbb{E}\left(\bar{X}_i\bar{Z}_i|X_i, Y_i\right)\right) \\
&= \mathbb{E}\left(\bar{X}_i\mathbb{E}\left(\bar{Z}_i|X_i, Y_i\right)\right) \\
&= \mathbb{E}\left(\bar{X}_i^2\right) \\
&= \text{Var}\left(Y_i\right) + \frac{\mathbb{E}\left(v_m(Y_i)\right)}{n_m} + \mathbb{E}\left(Y_i\right)^2 ,
\end{aligned} \tag{30}$$

where we have used (26) to expand the second moment of $\bar{X}_i$.

We can now use the five results (7), (10), (18), (29), and (30) to write the correlation given in (17) as:

$$\begin{aligned}\rho(\bar{X}_i, \bar{Z}_i) &= \frac{\mathbb{E}\left(\bar{X}_i\bar{Z}_i\right) - \mathbb{E}\left(\bar{X}_i\right)\mathbb{E}\left(\bar{Z}_i\right)}{\sqrt{\mathrm{Var}\left(\bar{X}_i\right)\mathrm{Var}\left(\bar{Z}_i\right)}} \\ &= \frac{\mathrm{Var}\left(Y_i\right) + \frac{1}{n_m}\mathbb{E}\left(v_m(Y_i)\right) + \mathbb{E}\left(Y_i\right)^2 - \mathbb{E}\left(Y_i\right)^2}{\sqrt{\left(\left(\frac{2n_m-1}{n_m^2}\right)\mathbb{E}\left(v_m(Y_i)\right) + \mathrm{Var}\left(Y_i\right)\right)\left(\mathrm{Var}\left(Y_i\right) + \frac{\mathbb{E}(v_m(Y_i))}{n_m}\right)}} \\ &= \sqrt{\frac{\mathrm{Var}\left(Y_i\right) + \frac{1}{n_m}\mathbb{E}\left(v_m(Y_i)\right)}{\mathrm{Var}\left(Y_i\right) + \left(\frac{2n_m-1}{n_m^2}\right)\mathbb{E}\left(v_m(Y_i)\right)}}\ .\end{aligned} \tag{31}$$

4.2 Mean-Squared Error

The squared-error between the resampled mean $\bar{Z}_i$ and the sample mean $\bar{X}_i$ is

$$\epsilon_i^2 = (\bar{Z}_i - \bar{X}_i)^2. \tag{32}$$

The mean-squared error (MSE), or D^2, is simply the mean of these squared differences,

$$D^2 = \frac{1}{n_f}\sum_{i=1}^{n_f}\epsilon_i^2. \tag{33}$$

This is a random variable, and we are seeking its expected value $\mathbb{E}\left(D^2\right)$.

$$\begin{aligned}\mathbb{E}\left(D^2\right) &= \mathbb{E}\left(\frac{1}{n_f}\sum_{i=1}^{n_f}\epsilon_i^2\right) = \frac{1}{n_f}\sum_{i=1}^{n_f}\mathbb{E}\left(\epsilon_i^2\right) \\ &= \mathbb{E}\left(\epsilon_i^2\right),\end{aligned} \tag{34}$$

where the final equality exploits the fact the ϵ_i^2 are i.i.d. Using the law of total expectation and the result in (28), we can show that

$$\begin{aligned}\mathbb{E}\left(\epsilon_i^2\right) &= \mathbb{E}\left((\bar{Z}_i - \bar{X}_i)^2\right) \\ &= \mathbb{E}\left(\mathbb{E}\left((\bar{Z}_i - \bar{X}_i)^2\right)|X_i, Y_i\right) \\ &= \mathbb{E}\left(\mathbb{E}\left((\bar{Z}_i - \mathbb{E}\left(\bar{Z}_i\right))^2\right)|X_i, Y_i\right) \\ &= \mathbb{E}\left(\mathrm{Var}\left(\bar{Z}_i|X_i, Y_i\right)\right) \\ &= \left(\frac{n_m - 1}{n_m^2}\right)\mathbb{E}\left(v_m(Y_i)\right).\end{aligned} \tag{35}$$

5. Discussion and Conclusion

We have derived relatively simple expressions for two correlations and two MSE's of interest in the situation outlined in Fig. 2.1. All four of these expressions contain only three quantities. First is the variance of the true values, and this remains fixed in all expressions, as one would expect. This variance could be described as the "signal" in this problem. The second quantity is the conditional expected value of the measurement variance function (the function that describes the variance in the measurements as a function of the true value of the measurand). This could be described as the "noise" in this problem. The final quantity is the number of measurements, and this value reduces the impact of the measurement variance, as one would expect when averaging noisy measurements. One could say that increasing the number of measurements reduces the measurement noise and thus improves the signal-to-noise ratio.

Equation (12) shows that the correlation between the sample mean and the true value is driven by the ratio of two variances. The numerator contains the variance of the true values. The denominator contains that same variance plus the expected value of the measurement variance function scaled down by the number of measurements. Variances are positive, so this ratio is clearly less than one. As the number of measurements increases, the measurement variance function becomes less important, the additive term goes to zero, and the ratio (the squared correlation) tends towards 1.0.

The top panel of Fig. 5.1 gives an example of (12) for the case where Var$(Y_i) = 4/3$ and $\mathbb{E}(v_m(Y)) = 1.0$. These same values are used for all examples that follow. Note that the specific distribution f_Y does not matter; only its variance matters. So, Var$(Y_i) = 4/3$ could result from the uniform case f_Y: Uniform$(0,1)$, or the normal case $f_Y : \mathcal{N}(0, 4/3)$, or any other distribution for f_Y. Similarly, the specific variance function $v_m(Y))$ does not matter; only its expected value matters. So, $\mathbb{E}(v_m(Y)) = 1.0$ could result from a variance that is constant or one that varies with Y in any number of different ways. Further note that due to the i.i.d. assumptions and use of population correlation, the number of files, n_f, does not appear in any of the four equations shown in the figure.

Equation (16) shows that the MSE between the sample mean and the true value is simply the expected value of the measurement variance function scaled down by the number of measurements. For a fixed number of measurements, a greater measurement variance function results in a greater MSE. And for any measurement variance function, as the number of measurements increases, the MSE tends towards zero. The bottom panel of Figure 5.1 provides an example of (16).

Equation (31) shows that the correlation between the sample mean and the resampled mean is driven by the ratio of two variances. As before, the variance of the true values appears in both the numerator and the denominator. In this case, the measurement variance function also appears in both, but with different scalings, each driven by the number of measurements. The scaling is such that the denominator is larger than the numerator for $n_m > 1$. And both scalings cause this term to vanish as the number of measurements grows, causing the correlation to tend towards 1.0. The top panel of Fig. 5.1 gives an example of (31).

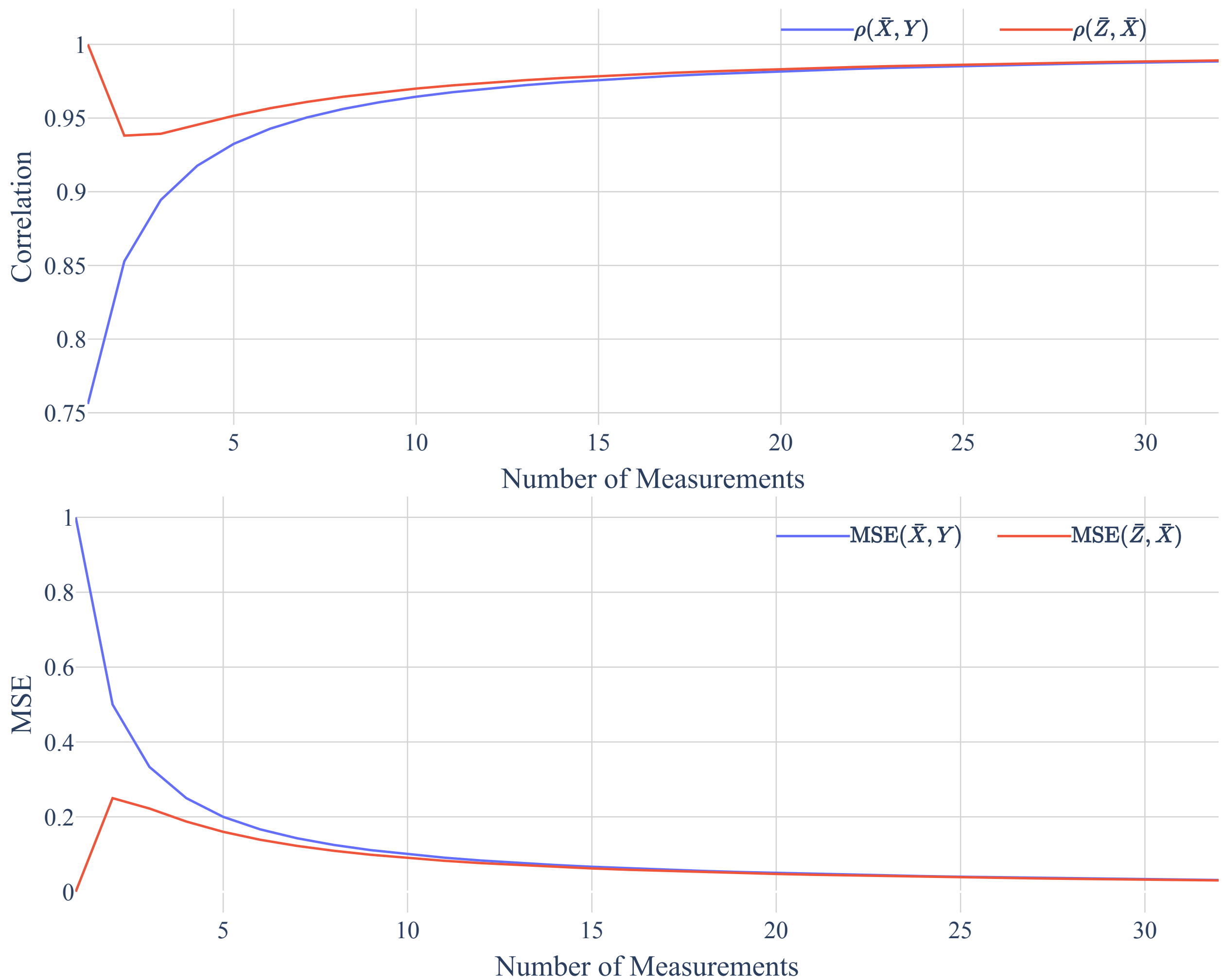

Figure 5.1. Example correlation and MSE results for 1 to 32 measurements for the case $Var(Y_i) = 4/3$ and $\mathbb{E}(v_m(Y)) = 1.0$. Blue lines show agreement between true value and sample mean. The top blue line is (12) and the bottom blue line is (16). Red lines show agreement between sample mean and resampled mean. The top red line is (31) and the bottom red line is (35).

Equation (35) shows that the MSE between the sample mean and resampled mean is simply the expected value of the measurement variance function scaled down by a function of number of measurements. For a fixed number of measurements, a greater measurement variance function results in a greater MSE. And for any measurement variance function, as the number of measurements increases, the MSE tends towards zero. The bottom panel of Fig. 5.1 provides an example of (35).

While these results agree with intuition, the exact dependence on the number of measurements is not necessarily evident at the outset. A fair amount of detailed mathematical derivation was required to produce them, and we have carried out simulations to verify that the results are indeed correct. Together, these results have provided building blocks for our additional work to better understand and harmonize labeled speech datasets where MOS labels may be based on a very wide range of individual votes (i.e., noisy measurements), leading to a corresponding range of correlations and MSEs.

Finally, we remark that the upper panel of Figure 5.1 suggests that if correlations near or above 0.97 are the goal, for example, then exceeding 10 measurements would achieve that goal. It is important to note that this result depends on the specific variances $\text{Var}(Y_i) = 4/3$ and $\mathbb{E}(v_m(Y)) = 1.0$ used in this example and one cannot draw general conclusions from a single example. In addition, the value of a ”suitably high” correlation will depend on the specifics of a given measurement problem and the value 0.97 given above is also just a single specific example. For any specific measurement problem, one must use the appropriate equations listed in this section to calculate PCC and MSE values for that problem and evaluate them in the context of that problem.

References

[1] International Telecommunication Union, Telecommunication Standardization Sector, Study Group 12, *ITU-T coded-speech database*, ITU T P Suppl. 23 (02/1998), ITU, Geneva, Switzerland, February 1998, https://handle.itu.int/11.1002/1000/4415

[2] A. Webster and F. Speranza, *Report on the validation of video quality models for high definition video content* (Video Quality Experts Group, June 2010), https://vqeg.org/media/4212/vqeg_hdtv_final_report_version_2.0.zip

[3] Chen, Yu-Wen, and Yu Tsao, "InQSS: A Speech Intelligibility and Quality Assessment Model Using a Multi-Task Learning Network," in *Proc. Interspeech 2022* (ISCA, September 2022), 3088–3092. https://doi.org/10.21437/Interspeech.2022-10153

[4] L. F. Tiotsop, A. Servetti, and E. Masala, "A Scoring Model Considering the Variability of Subjects' Characteristics in Subjective Experiments," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)* (IEEE, June 2023), 43–48. https://doi.org/10.1109/QoMEX58391.2023.10178476

[5] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application* (USA: Cambridge University Press, 2013). https://doi.org/10.1017/CBO9780511802843

REPORT DOCUMENTATION PAGE

1. REPORT DATE
March 2026

2. REPORT TYPE
Technical Memorandum

3. DATES COVERED

START DATE	END DATE
July 2026	February 2026

4. TITLE AND SUBTITLE
Analysis of Multiple Noisy Measurements and Their Resamplings via Expected Values of Correlation and Mean-Squared-Error

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER
3146012

5e. TASK NUMBER
300

5f. WORK UNIT NUMBER
NTIA/ITS.TAM

6. AUTHORS(S)
Jaden Pieper and Stephen D. Voran

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Telecommunication Sciences
National Telecommunications and Information Administration
U.S. Department of Commerce
325 Broadway
Boulder, CO 80305

8. PERFORMING ORGANIZATION REPORT NUMBER
NTIA TM-26-581

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Telecommunications and Information Administration
Herbert C. Hoover Building
1401 Constitution Ave., N.W.
Washington, D.C. 20230

10. SPONSOR/MONITOR'S ACRONYM(S)
NTIA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This memorandum provides mathematical relationships for agreement between true measurand values, means of multiple imperfect measurements, and resampled measurement means. Agreement is quantified here in terms of Pearson correlation coefficient (PCC) and mean-squared error (MSE). The derivations of the key relationships range from simple to rather intricate, but the resulting equations are all simple, intuitive, and easy to use. More specifically, given a number of measurands, a number of measurements per measurand, and some basic variance information, the equations provide the expected PCC (or MSE) between measurand values, means of measurements, and means of resampled measurements. Depending on the measurement context and the specific goals for PCC or MSE, these results can guide readers in deciding on an appropriate number of measurements to gather.

15. SUBJECT TERMS
bootstrapping, correlation, mean opinion score, mean-squared error, measurement error, MOS, resampling, subjective testing

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
Unclassified	Unclassified	Unclassified	UU	22

19a. NAME OF RESPONSIBLE PERSON
William Kozma Jr.

19b. PHONE NUMBER *(Include area code)*
303-335-5059