Why Mean, Variance, Moments, Correlation, Skewness etc. – Invariance-Based Explanations

Olga Kosheleva[1], Laxman Bokati[2], and Vladik Kreinovich[2,3]
[1]Department of Teacher Education
[2]Computational Science Program
[3]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, Texas 79968, USA
olgak@utep.edu, lbokati@miners.utep.edu, vladik@utep.edu

Abstract

In principle, we can use many different characteristics of a probability distribution. However, in practice, a few of such characteristics are mostly used: mean, variance, moments, correlation, etc. Why these characteristics and not others? The fact that these characteristics have been successfully used indicates that there must be some reason for their selection. In this paper, we show that the selection of these characteristics can be explained by the fact that these characteristics are *invariant* with respect to natural transformations – while other possible characteristics are *not* invariant.

1 Formulation of the Problem

Need for probabilistic models. One of the main objectives of science is *predict* the future state of the world, i.e., to predict the future values of the world's processes.

Some processes are *deterministic*. For example, in celestial mechanics, we can predict the locations of the planets hundreds of years from now – and indeed, such locations (and, in particular solar and lunar eclipses) – have been successfully predicted hundreds of years ago.

However, most other processes are *probabilistic*. We cannot predict the exact value of the stock market, we cannot predict tomorrow's temperature – but what we can usually predict reasonably well, based on our previous experiences, are *probabilities* of different future values.

Need for numerical characteristics of probabilistic models. In the computer, everything is stored as numbers. From this viewpoint, describing a future-

related probability distribution means describing certain numerical characteristics of this distribution.

If we consider probabilities describing the values of a single quantity, we need numerical characteristics of the corresponding 1-D probability distribution. If we consider probabilities describing the values of several quantities, we need numerical characteristic of the corresponding joint multi-D distribution.

Which characteristics should we select? In principle, there are many possible numerical characteristics of a probability distribution:

- we can use moments,
- we can use the values of the probability density function or of the cumulative distribution function,
- we can use the characteristic function of the distribution, etc.;

see, e.g., [6].

Which of these characteristics should we select?

Which characteristic are usually selected? Interestingly, in practice, only a few of these characteristics are routinely used.

If you give some raw 1-D data to a scientist or to an engineer, this scientist or engineer will first compute the mean and the standard deviation; maybe he or she will also compute the skewness. If you give them 2-D data they will also compute covariance and correlations. These numerical characteristic are so overwhelmingly used in practice that many scientific calculators have special buttons automatically computing these characteristics

But why? The fact that these characteristics have been actively used by practitioners means that indeed, in many practical situations, these particular characteristic have been very helpful. The fact that they have not been replaced by any other possible characteristics means that they are, in general, more helpful than others.

A natural question is: why are these characteristics more helpful than others?

What we do in this paper. In this paper, we provide an answer to the above "why" question. Namely, we show that:

- the most widely used numerical characteristic of probability distributions are *invariant* with respect to natural transformations, while
- other possible characteristics are *not* invariant.

This explains why the selected characteristics are used.

Structure of this paper. We start, in Section 2, with analyzing what are possible numerical characteristic of probability distributions. Then, in Section 3, we describe natural symmetries and corresponding invariances. In Section 4, we formulate the main result: that only moments – and characteristic determined

by different moments – are invariant. In Section 5, we explain the ubiquity of specific combinations of moments such as variance, correlation, and skewness. For readers' convenience, all the proofs are placed in a special Proofs Section 6.

2 Towards a General Description of Possible Numerical Characteristics of Probability Distributions

Need for decision making. The ultimate goal of predictions is to make decisions. If we know where the stock market will go, we should either buy or sell the corresponding stocks. If we know tomorrow's temperature, then we should dress accordingly – and, if needed, get prepared to protect the plants against a sudden cold.

So, when we select what numerical characteristics of probability distributions, we should take into account that these characteristics must be useful for making a decision. In order to make a good decision, we need to have a good understanding of the person's preferences. Let us briefly recall how these preferences are usually described and how we can make a decision based on these preferences; for a detailed description, see, e.g., [1, 2, 3, 4, 5].

How can we describe human preferences. In order to describe a person's preferences, a reasonable idea is to select two extreme alternatives, more extreme that anything that we will actually encounter:

- a very good alternative A_+ which is better than anything that we will actually encounter, and
- a very bad alternative A_- which is worse than anything that we will actually encounter.

Then, for each number p from the interval $[0, 1]$, we can form a lottery – that we will denote by $L(p)$ – in which:

- we get A_+ with probability p, and
- we get A_- with the remaining probability $1 - p$.

When $p = 0$, we have $L(0) = A_-$, so the corresponding lottery is worse than any actual alternative A; we will denote this by $A_- < A$. As the probability p increases, the lottery becomes better and better, and for $p = 1$, we have $L(1) = A_+$ and thus, $A < A_+$.

It is easy to show that there exists a threshold $\sup\{p : L(p) < A\} = \inf\{p : A < L(p)\}$ that separates probabilities for which A is better from probabilities for which the lottery is better. This threshold value is known as the *utility* of the alternative A. It is usually denoted by $u(A)$.

By definition of utility, for any small value $\varepsilon > 0$, we have

$$L(u(A) - \varepsilon) < A < L(u(A) + \varepsilon).$$

For very small values ε, the difference between the probabilities $u(A)$, $u(A)-\varepsilon$, and $u(A)+\varepsilon$ is practically indistinguishable. In this sense, we can say that the alternative A is *equivalent* to the lottery $L(u(A))$. We will denote this equivalence by $A \equiv L(u(A))$.

Clearly, if $p < p'$, this means that the lottery $L(p')$ is better. Thus, if $u(A) < u(B)$, we have $L(u(A)) < L(u(B))$ and, since $A \equiv L(u(A))$ and $B \equiv L(u(B))$, that $A < B$. So, one alternative is better than the other if its utility is larger.

How can we make a decision? In practice, when we make a decision, we do not know the exact consequence of each of the possible actions a. At best, we can, based on our prior experiences, estimate the probabilities $p_1, \ldots, p_n$ of possible consequences $A_1, \ldots, A_n$. Let $u_i \stackrel{\text{def}}{=} u(A_i)$ denote the utility of the i-th alternative.

Each alternative A_i is equivalent to the corresponding lottery $L(u_i)$. Thus, for the decision maker, the consequences of selecting an action a are equivalent to a two-stage lottery, in which:

- first, we select one of the consequences A_i with probability p_i, and then,
- depending on which consequence A_i we selected on the first stage, we select the very good alterative A_+ with probability u_i and the very bad alternative A_- with probability $1-u_i$.

As a result of this two-stage lottery, we end up either with A_+ or with A_-, and the probability of selecting A_+ is equal to

$$p_1 \cdot u_1 + \ldots + p_n \cdot u_n.$$

By definition, this probability is the utility $u(a)$ of selecting an action a. Thus, this utility is equal to the above expression:

$$u(a) = p_1 \cdot u_1 + \ldots + p_n \cdot u_n.$$

We want to select the best action, i.e., the action with the largest possible value of utility. In mathematical terms, the above formula for the utility of the action simply means that the action's utility is equal to the expected value $E[u_i]$ of the utility. So, to make a proper decision, we need to know expected values $E[u(x)]$ of different functions $u(x)$ – namely, functions describing the person's utility. Here, x may be a single parameter, may be several parameters.

Usually, small changes in x lead to equally small changes in our utility: e.g., we do not expect much difference between temperatures 24 C or 25 C, or between predicting that the Dow-Jones will rise by 101 or by 102 points. Thus, it is reasonable to require that the utility function $u(x)$ is smooth (= differentiable). Thus, we arrive at a following definition.

Definition 2.1. *Let $n \geq 1$ by an integer. By a* characteristic*, we mean a mapping that assigns, to each random vector $X = (X_1, \ldots, X_n)$, a value $E[f(X_1, \ldots, X_n)]$, where $f(x_1, \ldots, x_n)$ is a smooth function of n variables.*

Comment. According to our definition, characteristics and in 1-1 correspondence with smooth functions. Thus, to make the exposition clearer, in the following text, we will sometimes identify a characteristic with the corresponding function $f(x_1, \ldots, x_n)$.

Examples.

- For $n = 1$ and $f(x_1) = x_1$, we get the mean.
- For $n = 1$ and $f(x_1) = x_1^2$, we get the second moment.
- For $n = 1$ and $f(x_1) = \exp(\omega \cdot x_1 \cdot \mathrm{i})$, we get different values of the characteristic function, etc.

Need to select a finite set of characteristics. In the computer, we can store only finitely many numbers. Thus, we need to select a finite set of characteristics.

Some sets are equivalent: e.g., if we know the mean and the second moment, then we can also compute the expected value of the functions $2x$ and $2x^2$, and vice versa. Let us describe a general definition.

Definition 2.2. *We say that the set of characteristics $\{f_1, \ldots, f_m\}$ and $\{g_1, \ldots, g_p\}$ are* equivalent *if the following two conditions are satisfied:*

- *the values $E[f_1]$, ..., $E[f_m]$ of the characteristics from the first set uniquely determine the values of all the characteristics $E[g_1]$, ..., $E[g_p]$ from the second set, and*
- *the values $E[g_1]$, ..., $E[g_p]$ of the characteristics from the second set uniquely determine the values of all the characteristics $E[f_1]$, ..., $E[f_m]$ from the first set.*

Proposition 2.1. *The two sets of characteristics $\{f_1, \ldots, f_m\}$ and $\{g_1, \ldots, g_p\}$ are equivalent if and only if the following two conditions are satisfied:*

- *each function $g_j(x)$ from the second set is equal to a linear combination of functions from the first set and 1, i.e., if there exist coefficients a_{ji} for which, for all j and all x, we have*

$$g_j(x) = a_{j0} + a_{j1} \cdot f_1(x) + \ldots + a_{jm} \cdot f_m(x); \text{ and}$$

- *each function $f_i(x)$ from the second set is equal to a linear combination of functions from the first set and 1, i.e., if there exist coefficients b_{ij} for which, for all i and all x, we have*

$$f_i(x) = b_{i0} + b_{i1} \cdot g_1(x) + \ldots + b_{ip} \cdot g_p(x).$$

3 Natural Symmetries and Corresponding Invariances

Possibility of re-scaling. In data processing, we process the numerical values of different quantities. It is important to mention, however, that for exact same state of the world, the corresponding numerical values will change if we change the measuring unit. For example:

- If we measure distances in km and then decide to switch to meters, then all the numerical values will multiply by 1000.
- If, in our borderline region between the US and Mexico, we change the monetary units from US dollars to Mexican pesos, then all the numerical values are multiplied by approximately 20 (or whatever the exchange rate will be).

In general, if we replace the original measuring unit with a $\lambda > 0$ times smaller one, then all the numerical values will be multiplied by λ: $x \to \lambda \cdot x$. This transformation is known as *re-scaling*.

Comment. In the above paragraph, we explained re-scaling corresponding to positive values λ. In some situations, negative values are also possible. For example:

- For the electric charge (and for the related quantities such as electric current), the sign has been rather arbitrarily chosen. Nothing will change if we view what was previously considered positive as negative and vice versa.
- In economics, the positive trade deficit in a trade of country A with country B is equivalent to a negative deficit when considered from the viewpoint of country B.

In view of this possibility, in the following text, we will consider re-scalings with negative coefficients λ as well.

Need for scale-invariance. Since the selection of a measuring unit is usually rather arbitrary, it makes sense to require that the result of data processing not depend on the choice of the measuring unit, i.e., that we should come up with the same conclusion if we start with re-scaled data.

Possibility of shift. For many quantities, the numerical value also depends on the starting point. For example:

- when we measure time, we can start from Year 0, or we can start with the beginning of the financial year, or with the beginning of the quarter;
- when we measure temperature, we can start with the temperature at which water freezes – as in Celsius scale – or with another starting point as, e.g., in the Fahrenheit scale;

- when we estimate the country's average or median income, we can consider the absolute income – or, which makes some sense, we can subtract, from each income, the minimum necessary to maintain living, and only compare values in excess of this minimum.

In general, if we replace the original starting point with a new starting point which is c unit before it, then this number c will be added to all the numerical values $x \to x + c$. This transformation is known as *shift*.

Need for shift-invariance. Since the selection of a starting point is often rather arbitrary, it makes sense to require that the result of data processing not depend on the choice of the starting point, i.e., that we should come up with the same conclusion if we start with shifted data.

4 Invariant Characteristics: This Explains Why Moments

Let us apply invariance ideas to selection of characteristics. In view of the arguments presented in the previous section, it is desirable to select characteristics in such a way that the resulting information not change if we re-scale or shift the numerical values.

Definition 4.1. *We say that a finite set of characteristics*

$$\{f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)\}$$

is shift-invariant *if for every tuple $c = (c_1, \ldots, c_n)$, once we know the values*

$$E[f_1(X_1, \ldots, X_n)], \ldots, E[f_m(X_1, \ldots, X_n)],$$

then we should be able to uniquely determine the values

$$E[f_i(X_1 + c_1, \ldots, X_n + c_n)]$$

for all i.

Definition 4.2. *We say that a finite set of characteristics*

$$\{f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)\}$$

is scale-invariant *if for every tuple $c = (c_1, \ldots, c_n)$, once we know the values*

$$E[f_1(X_1, \ldots, X_n)], \ldots, E[f_m(X_1, \ldots, X_n)],$$

then we should be able to uniquely determine the values $E[f_i(c_1 \cdot X_1, \ldots, c_n \cdot X_n)]$ for all i.

Discussion. To describe all possible shift- and scale-invariant sets of characteristics, we need to introduce the following auxiliary definitions.

Definition 4.3. *By a* moment*, we mean a characteristic corresponding to* $f(x_1, \ldots, x_n) = x_1^{k_1} \cdot \ldots \cdot x_n^{k_n}$*, for some non-negative integers* k_i*.*

Notation. In the following text, the corresponding values $E[f]$ with be denoted by letter M with indices listing each variable k_i times. For example, $E[X_i]$ will be denoted by M_i, $E\left[X_i^2\right]$ by M_{ii}, $E[X_i \cdot X_j]$ by M_{ij}, etc.

Definition 4.4. *We say that a finite set of moments is an* ideal of moments *if for each moment* $x_1^{k_1} \cdot \ldots \cdot x_n^{k_n}$*, this set also includes all the moments* $x_1^{k_1'} \cdot \ldots \cdot x_n^{k_n'}$ *for which* $k_i' \le k_i$ *for all* i*.*

Examples.

- All first moments $M_1, \ldots, M_n$ form an ideal.
- The set of all first and second moments M_i and M_{ij} forms an ideal, etc.

Discussion. Now, we are ready to formulate our main result.

Proposition 4.1. *For each finite set of characteristics* $\{f_1, \ldots, f_m\}$*, the following two conditions are equivalent to each other:*

- *the set of characteristics is shift- and scale-invariant, and*
- *the set of characteristics is equivalent to an ideal of moments.*

Discussion. This results explains the ubiquity of moments.

5 Invariant Combinations of Characteristics: This Explains Why Variance, Covariance, Coefficient of Variation, Correlation, and Skewness

What we do in this section. In the previous section, we showed that with respect to natural transformations, the only invariant characteristics are, in effect, moments.

The next question is why some combinations of moments are actively used – while others are used rarely. In this section, we show this can also be explained by invariance. Specifically, we show that invariances explains the ubiquity of five such widely used combinations: variance, covariance, correlation, coefficient of variation, and skewness.

Comment. In contrast to a new (and not so easy to prove) result from the previous section, results from this section are largely known – and are easy to prove. We included these results into the paper, since they nicely supplement the explanation provided in the previous section – of why moments and their combinations are mostly used – by explaining the ubiquity of several specific combinations of moments.

Definition 5.1. *We say that a mapping $F(X_1, \ldots, X_n)$ that assigns a numerical value to each random vector $(X_1, \ldots, X_n)$ is* shift-invariant *if for each random vector $X = (X_1, \ldots, X_n)$ and each tuple $c = (c_1, \ldots, c_n)$ of real numbers, F assigns the same value to the original random vector $(X_1, \ldots, X_n)$ and to its shift $(X_1 + c_1, \ldots, X_n + c_n)$:*

$$F(X_1, \ldots, X_n) = F(X_1 + c_1, \ldots, X_n + c_n).$$

Definition 5.2. *We say that a mapping $F(X_1, \ldots, X_n)$ that assigns a numerical value to each random vector $(X_1, \ldots, X_n)$ is* scale-invariant *if for each random vector $X = (X_1, \ldots, X_n)$ and each tuple $c = (c_1, \ldots, c_n)$ of real numbers, F assigns the same value to the original random vector $(X_1, \ldots, X_n)$ and to its re-scaling $(c_1 \cdot X_1, \ldots, c_n \cdot X_n)$:*

$$F(X_1, \ldots, X_n) = F(c_1 \cdot X_1, \ldots, c_n \cdot X_n).$$

Discussion. Which combinations of moments are shift and/or scale-invariant? Let us first consider combinations of first moments $M_i = E[X_i]$.

Proposition 5.1. *No combination $f(M_1, \ldots, M_n)$ of first order moments is shift-invariant.*

Proposition 5.2. *No combination $f(M_1, \ldots, M_n)$ of first order moments is scale-invariant.*

Discussion. If we also allow second-order moments $M_{ij} = E[X_i \cdot X_j]$, then shift- and/or scale-invariant combinations become possible.

Proposition 5.3. *A combination $f(\{M_i\}, \{M_{ij}\})$ of the first two moments is shift-invariant if and only if it a function of the variances $V_i = M_{ii} - M_i^2$ and covariances $C_{ij} = M_{ij} - M_i \cdot M_j$.*

Discussion. This result explains the ubiquity of variance and covariance.

Proposition 5.4. *A combination $f(\{M_i\}, \{M_{ij}\})$ of the first two moments is scale-invariant if and only if it a function of the coefficients of variation $CV_i = \frac{\sigma_i}{M_i}$ (where $\sigma_i \stackrel{\text{def}}{=} \sqrt{V_i}$) and the coefficients of covariance $CV_{ij} = \frac{C_{ij}}{M_i \cdot M_j}$.*

Discussion.

- This result explains the ubiquity of the coefficient of variation.
- For second-order moments, it is also possible to have combinations which are both shift- and scale-invariance.

Proposition 5.5. *A combination $f(\{M_i\}, \{M_{ij}\})$ of the first two moments is shift- and scale-invariant if and only if it a function of the correlations*

$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \cdot \sigma_j}.$$

Discussion.

- This result explains the ubiquity of correlation.
- For the case when we have only one variable, the above result shows that no combination of the first and second moments is shift- and scale-invariant. It turns out that such an invariant combination is possible if we also allow the third moment.

Proposition 5.6. *A combination $f(M_1, M_{11}, M_{111})$ of the first three moments is shift- and scale-invariant if and only if it a function of the skewness*

$$\widetilde{\mu}_3 = E\left[\left(\frac{X_1 - M_1}{\sigma_1}\right)^3\right].$$

Discussion.

- This result explains the ubiquity of skewness.
- If we also allow the fourth moment, we get a function of skewness and kurtosis:

Proposition 5.7. *A combination $f(M_1, M_{11}, M_{111}, M_{1111})$ of the first four moments is shift- and scale-invariant if and only if it a function of the skewness $\widetilde{\mu}_3$ and of the kurtosis $\widetilde{\mu}_4 = E\left[\left(\frac{X_1 - M_1}{\sigma_1}\right)^4\right]$.*

6 Proofs

Proof of Proposition 2.1. Clearly, if

$$g_j(x) = a_{j0} + a_{j1} \cdot f_1(x) + \ldots + a_{jm} \cdot f_m(x),$$

then

$$E[g_j] = a_{j0} + a_{j1} \cdot E[f_1] + \ldots + a_{jm} \cdot E[f_m].$$

So, if we know the values $E[f_i]$, we will indeed be able to uniquely determine the values of $E[g_j]$ – and vice versa.

Thus, to prove the proposition, it is sufficient to prove that if, e.g., a function $g_1(x)$ cannot be represented as the desired linear combination, then we cannot

uniquely determine the value of $E[g_1]$ based on the known values of $E[f_i]$. Indeed, let us assume that g_1 is not equal to a linear combination of the functions f_i and 1. On the space of all the functions, we have a natural scalar (dot) product $\langle f, g\rangle \stackrel{\text{def}}{=} \int f(x) \cdot g(x)\, dx$.

We can then use the usual Gram-Schmidt orthonormalization in the linear space spanned by the functions f_i and 1 and find, as their linear combinations. the orthonormal vectors $e_1, \ldots, e_q$ that span the exact same linear space – and for which $\langle e_i, e_i\rangle = 1$ for all i and $\langle e_i, e_j\rangle = 0$ for all $i \neq j$. Then, the function $g_1(x)$ can be represented as

$$g_1(x) = \langle g_1, e_1\rangle \cdot e_1(x) + \ldots + \langle g_1, e_q\rangle \cdot e_q(x) + e(x),$$

where the difference

$$e(x) \stackrel{\text{def}}{=} g_1(x) - \langle g_1, e_1\rangle \cdot e_1(x) - \ldots - \langle g_1, e_q\rangle \cdot e_q(x)$$

is:

- orthogonal to all the vectors $e_i(x)$ – and thus, to their linear combinations f_i and 1, and
- different from 0 – since otherwise, $g_1(x)$ would be equal to a linear combination of the functions f_i and 1.

Due to the fact that $g_1(x)$ is orthogonal to all the functions $e_i(x)$, we conclude that $\langle g_1, e\rangle = \langle e, e\rangle$ and, since the difference $e(x)$ is not 0, we have

$$\langle g_1, e\rangle = \langle e, e\rangle > 0.$$

Let us now take a probability distribution which is everywhere positive on some interval – e.g., a uniform distribution, with the probability density function $\rho(x) = \text{const}$. Then, for small ε, the function $\rho_1(x) \stackrel{\text{def}}{=} \rho(x) + \varepsilon \cdot e(x)$ is also everywhere positive. Since the function $e(x)$ is orthogonal to 1, i.e., $\int e(x)\, dx = 0$, we get $\int \rho_1(x)\, dx = \int \rho(x)\, dx = 1$, so $\rho_1(x)$ is also a probability distribution. Since $e(x)$ is orthogonal to all the functions $f_i(x)$, we have

$$E_1[f_i] = \int f_i(x) \cdot \rho_1(x)\, dx = \int f_i(x) \cdot \rho(x)\, dx = E[f_i]$$

for all i. On the other hand,

$$E_1[g_1] = \int g_1(x) \cdot \rho_1(x)\, dx = \int g_1(x) \cdot \rho(x)\, dx + \varepsilon \cdot \int g_1(x) \cdot e(x)\, dx =$$

$$E[g_1] + \varepsilon \cdot \int g_1(x) \cdot e(x)\, dx.$$

We know that $\int g_1(x) \cdot e(x)\, dx = \langle g_1, e\rangle \neq 0$, so $E_1[g_1] \neq E[g_1]$.

Thus, we have two distributions ρ and ρ_1 for which the expected values $E[f_i]$ are the same, but the expected values of $E[g_1]$ are different. Thus, the sets $\{f_i\}$ and $\{g_j\}$ are indeed not equivalent. The proposition is proven.

Proof of Proposition 4.1.

1°. One can easily check that the set of characteristics corresponding to an ideal of moments is shift- and scale-invariant – and thus, that each equivalent set of characteristics is also shift- and scale-invariant. So, to complete the proof, we need to show that every shift- and scale-invariant set of characteristics is equivalent to an ideal of moments.

2°. Let us first analyze the consequences of shift-invariance. Due to Proposition 2.1, shift-invariance implies that for all possible shifts $c = (c_1, \ldots, c_n)$, each shifted function $f_i(x_1 + c_1, \ldots, x_n + c_n)$ is equal to a linear combination of the original functions and of 1, with coefficients possibly depending on c_i:

$$f_i(x_1 + c_1, \ldots, c_n + c_n) = a_{i0}(c_1, \ldots, c_n) +$$

$$a_{i1}(c_1, \ldots, c_n) \cdot f_1(x_1, \ldots, x_n) + \ldots + a_{im}(c_1, \ldots, c_n) \cdot f_m(x_1, \ldots, x_n). \quad (1)$$

Let us first consider the shift by one of the variables. Without losing generality, we will assume that this variable is x_1. Let us fix the values $x_2^{(0)}, \ldots, x_n^{(0)}$ of all other variables, i.e., let us consider functions of one variable $F_i(x_1) = f_i\left(x_1, x_2^{(0)}, \ldots, x_n^{(0)}\right)$ and $A_{ij}(c_1) = a_{ij}(c_1, 0, \ldots, 0)$. For these functions, the above equation takes a simplified form

$$F_i(x_1 + c_1) = A_{i0}(c_1) + A_{i1}(c_1) \cdot F_1(x_1) + \ldots + A_{im}(c_1) \cdot F_m(x_1). \quad (2)$$

In this equality, for each i, we have $m + 1$ unknown functions $A_{ij}(c_1)$. To find the values of these functions, let us select $m + 1$ different value of x_1:

$$x_1^{(0)}, \ldots, x_1^{(m)}.$$

Substituting these $m+1$ values into the formula (2), we get the following system of $m + 1$ linear equations for $m + 1$ unknowns $A_{ij}(c_1)$:

$$F_i\left(x_1^{(0)} + c_1\right) = A_{i0}(c_1) + A_{i1}(c_1) \cdot F_1\left(x_1^{(0)}\right) + \ldots + A_{im}(c_1) \cdot F_m\left(x_1^{(0)}\right);$$

$$\ldots$$

$$F_i\left(x_1^{(m)} + c_1\right) = A_{i0}(c_1) + A_{i1}(c_1) \cdot F_1\left(x_1^{(m)}\right) + \ldots + A_{im}(c_1) \cdot F_m\left(x_1^{(m)}\right).$$

By Cramer's rule, the solution to this system is a linear combination of the right-hand sides $F_i\left(x_1^{(k)} + c_1\right)$ with coefficients depending on the values $F_j\left(x_1^{(k)}\right)$ and thus, not depending on c_1. The functions F_i are smooth, thus their linear combination is also smooth. So, all the functions $A_{ij}(c_1)$ are differentiable.

Since all the functions in the equality (2) are differentiable, we can differentiate both sides with respect to c_1 and then take $c_1 = 0$. As a result, we get

$$F_i'(x_1) = \alpha_{i0} + \alpha_{i1} \cdot F_1(x_1) + \ldots + \alpha_{im} \cdot F_m(x_1),$$

where we denoted $\alpha_{ij} \stackrel{\text{def}}{=} A'_{ij}(0)$. We have such an equation for each i. Thus, for m unknown functions $F_i(x_1)$, we have a system of linear differential equations with constant coefficients:

$$F'_1(x_1) = \alpha_{10} + \alpha_{11} \cdot F_1(x_1) + \ldots + \alpha_{1m} \cdot F_m(x_1);$$

$$\ldots$$

$$F'_m(x_1) = \alpha_{m0} + \alpha_{n1} \cdot F_1(x_1) + \ldots + \alpha_{mm} \cdot F_m(x_1).$$

We can transform this system to a more standard form if we add an auxiliary function $F_0(x_1) = 1$ with equation $F'_0(x_1) = 0$ and replace α_{i0} with an equivalent expression $\alpha_{i0} \cdot F_0(x_1)$:

$$F'_0(x_1) = 0;$$

$$F'_1(x_1) = \alpha_{10} \cdot F_0(x_1) + \alpha_{11} \cdot F_1(x_1) + \ldots + \alpha_{1m} \cdot F_m(x_1);$$

$$\ldots$$

$$F'_m(x_1) = \alpha_{m0} \cdot F_0(x_1) + \alpha_{n1} \cdot F_1(x_1) + \ldots + \alpha_{mm} \cdot F_m(x_1).$$

It is known that a general solution to such a system of equations is a linear combination of functions $x_1^k \cdot \exp(z \cdot x_1)$, where z are eigenvalues of the matrix α_{ij}, and a natural number k does not exceed the multiplicity of the corresponding eigenvalue minus 1 – i.e., in this case, $k \le m$. In general, eigenvalues are complex numbers $z = a + b \cdot \mathrm{i}$. In terms of real numbers, the general solution is a linear combination of the functions $x_1^k \cdot \exp(a \cdot x_1) \cdot \sin(b \cdot x_1)$ and $x_1^k \cdot \exp(a \cdot x_1) \cdot \cos(b \cdot x_1)$.

3°. Let us now consider the consequences of scale-invariance. Similar to Part 2 of this proof, we get the formula

$$F_i(c_1 \cdot x_1) = B_{i0}(c_1) + B_{i1}(c_1) \cdot F_1(x_1) + \ldots + B_{im}(c_1) \cdot F_m(x_1) \qquad (3)$$

for some functions $B_{ij}(c_1)$. Similarly to Part 2, we can conclude that the functions $B_{ij}(c_1)$ are smooth. Thus, we can differentiate both sides of the formula (3) with respect to c_1, take $c_1 = 1$, and thus, get the following equation:

$$x_1 \cdot F'_i(x_1) = \beta_{i0} + \beta_{i1} \cdot F_1(x_1) + \ldots + \beta_{im} \cdot F_m(x_1),$$

where we denoted $\beta ij \stackrel{\text{def}}{=} B'_{ij}(0)$. The left-hand side can be rewritten as

$$x_1 \cdot \frac{dF_i}{dx_1} = \frac{dF_1}{dx_1/x_1},$$

i.e., as $\dfrac{dF_1}{dL}$, where we denoted $L \stackrel{\text{def}}{=} \ln(x_1)$ (so that $x_1 = \exp(L)$). Hence, in terms of the new variable L, for the corresponding functions $G_i(L) = F_i(\exp(L))$, we get

$$G'_i(L) = \beta_{i0} + \beta_{i1} \cdot G_1(L) + \ldots + \beta_{im} \cdot G_m(L).$$

With respect to L, we again get a system of linear differential equations with constant coefficients. So, its general solution is a linear combination of the functions $L^k \cdot \exp(a \cdot L) \cdot \sin(b \cdot L)$ and $L^k \cdot \exp(a \cdot L) \cdot \cos(b \cdot L)$. Substituting $L = \ln(x_1)$ into these formulas and taking into account that $\exp(a \cdot \ln(x_1)) = (\exp(\ln(x_1))^a = x_1^a$, we conclude that $F_1(x_1)$ is a linear combination of functions $(\ln(x_1))^k \cdot x_1^a \cdot \sin(b \cdot \ln(x_1))$ and $(\ln(x_1))^k \cdot x_1^a \cdot \cos(b \cdot \ln(x_1))$.

4°. From Part 2 and 3, we see each function $F_i(x_1)$ has to be represented in two different forms. One can show that the only expression common to both forms is x_1^k for some natural $k \le m$. Thus, each function $F_i(x_1)$ is a linear combination of such expressions – and is, thus, a polynomial – and a polynomial of order $\le m$.

5°. Now we know for each combination of $x_2, \ldots$, the dependence on x_1 is a polynomial of order $\le m$. The coefficients of this polynomial, in general, depend on the values $x_2, x_3, \ldots$ So, if we fix the values $x_3^{(0)}, \ldots, x_n^{(0)}$, then for the corresponding function $H_i(x_1, x_2) = f_i\left(x_1, x_2, x_3^{(0)}, \ldots, x_n^{(0)}\right)$, we have

$$H_i(x_1, x_2) = a_0(x_2) + a_1(x_2) \cdot x_1 + \ldots a_m(x_2) \cdot x_1^m. \tag{4}$$

Similarly, when we fix x_1, the dependence on x_2 can also be described as a polynomial of degree $\le m$:

$$H_i(x_1, x_2) = b_0(x_1) + b_1(x_1) \cdot x_2 + \ldots b_m(x_1) \cdot x_2^m,$$

so

$$a_0(x_2) + a_1(x_2) \cdot x_1 + \ldots a_m(x_2) \cdot x_1^m = b_0(x_1) + b_1(x_1) \cdot x_2 + \ldots + b_m(x_1) \cdot x_2^m. \tag{5}$$

To determine $m+1$ coefficients $a_i(x_2)$, let us select $m+1$ different value of x_1: $x_1^{(0)}, \ldots, x_1^{(m)}$. Substituting these $m+1$ values into the formula (5), we get a system of $m+1$ linear equations for $m+1$ unknowns $a_i(x_2)$:

$$a_0(x_2) + a_1(x_2) \cdot x_1^{(0)} + \ldots a_m(x_2) \cdot \left(x_1^{(0)}\right)^m =$$

$$b_0\left(x_1^{(0)}\right) + b_1\left(x_1^{(0)}\right) \cdot x_2 + \ldots + b_m\left(x_1^{(0)}\right) \cdot x_2^m;$$

$$\ldots$$

$$a_0(x_2) + a_1(x_2) \cdot x_1^{(m)} + \ldots a_m(x_2) \cdot \left(x_1^{(m)}\right)^m =$$

$$b_0\left(x_1^{(m)}\right) + b_1\left(x_1^{(m)}\right) \cdot x_2 + \ldots + b_m\left(x_1^{(m)}\right) \cdot x_2^m.$$

By Cramer's rule, the solution to this system is a linear combination of the right-hand sides – which are polynomials in x_2 – with coefficients depending on the values $\left(x_1^{(m)}\right)^k$ (and thus, not depending on x_2). A linear combination of polynomials is also a polynomial. So, all the coefficients $a_i(x_2)$ are polynomials and thus, the expression (4) is a polynomial of two variables x_1 and x_2.

Similarly, we can prove that it is a polynomial of x_1, x_2, and x_3, etc., until we prove that each original function $f_i(x_1, \ldots, x_n)$ is a polynomial.

6°. To complete the proof, we must show that the corresponding set of polynomials is equivalent to an ideal of moments. Indeed, let us show that it is equivalent to the set of all monomials $x_1^{k_1} \cdot \ldots \cdot x_n^{k_n}$ that are parts of the polynomials f_i, plus monomials with $k_i' \le k_i$ for all i.

Of course, if we have all these monomials, then we can get all the polynomials f_i as their linear combinations. So, the only thing we need to prove is that if we know the value of $E[f_i]$, then we know the values $E[m]$ for all monomials forming f_i – as well as for all monomials with $k_i' \le k_i$. Let us perform this "separation" variable by variable. Let us start with the variable x_1. In terms of x_1 the polynomial f_i can be represented as

$$f_i = a_0 + a_1 + \ldots + a_m,$$

where a_k combines terms proportional to x_1^k. For each such term,

$$a_k(c_1 \cdot x_1, x_2, \ldots, x_m) = c_1^k \cdot a_k(x_1, \ldots, x_k).$$

Due to scale-invariance, for each c_1, the function

$$f_i(c_1 \cdot x_1, x_2, \ldots, x_n) = a_0 + c_1 \cdot a_1 + \ldots + c_1^m \cdot a_m \tag{5}$$

is a linear combination of the original functions $f_1, \ldots, f_m$.

We can select $m+1$ different values c_1: $c_1^{(0)}, \ldots, c_1^{(m)}$. Substituting these values into the formula (5), we get a system of $m+1$ linear equations with constant coefficients for $m+1$ unknowns a_i:

$$f_i\left(c_1^{(0)} \cdot x_1, x_2, \ldots, x_n\right) = a_0 + c_1^{(0)} \cdot a_1 + \ldots + \left(c_1^{(0)}\right)^m \cdot a_m;$$

$$\ldots$$

$$f_i\left(c_1^{(m)} \cdot x_1, x_2, \ldots, x_n\right) = a_0 + c_1^{(m)} \cdot a_1 + \ldots + \left(c_1^{(m)}\right)^m \cdot a_m.$$

A general solution to this system is a linear combination of the left-hand sides. Since each left-hand side is a linear combination of the original functions f_j, we conclude that all the functions a_i are also linear combinations of the original functions f_j.

Each function a_i has x_1 only in one power. Similarly, we can "split" each expression a_i into sub-expressions corresponding to different powers of x_2, etc. – until we conclude that all the monomials from each original polynomial can be represented as linear combinations of the original functions f_j.

The last thing we need to prove is that if we have a monomial $x_1^{k_1} \cdot \ldots \cdot x_n^{k_n}$, then for each $k_i' \le k_i$, we also have a monomial $x_1^{k_1'} \cdot \ldots \cdot x_n^{k_n'}$. Indeed, due to shift-invariance, with the original monomial $x_1^{k_1} \cdot \ldots \cdot x_n^{k_n}$, the shifted function $(x_1+1)^{k_1} \cdot \ldots \cdot (x_n+1)^{k_n}$ is also a linear combination of the original polynomials f_j. The expansion of this function into monomials includes all the

monomials $x_1^{k'_1} \cdot \ldots \cdot x_n^{k'_n}$ with $k'_i \le k_i$. So, as we have proved earlier, all these monomials with $k'_i \le k_i$ are also linear combinations of the original functions f_j.

The equivalence between the original set and the ideal of moments is thus proven, and so is the proposition.

Proof of Proposition 5.1: it follows from the Proposition 5.3 (see below).

Proof of Proposition 5.2: it follows from the Proposition 5.4 (see below).

Proof of Proposition 5.3. It is easy to check that variances and covariances are shift-invariant, and thus, that any combination of variances and covariances is also shift-invariant.

Let us prove that, vice versa, any shift-invariant combination is a function of variances and covariances. Indeed, by definition of shift-invariance, the value of this combination should not change if we shift the original random vector. In particular, we can shift it by subtracting the means, i.e., by taking $c_i = -M_i$. Then, for the shifted random variable $X'_i = X_i - M_i$, the first moments M'_i will be equal to 0. For the second moments, we have

$$M'_{ij} = E[X'_i \cdot X'_j] = E[(X_i - M_i) \cdot (X_j - M_j)] =$$

$$E[X_i \cdot X_j - X_i \cdot M_j - M_i \cdot X_j + M_i \cdot M_j] = E[X_i \cdot X_j] - M_j \cdot E[X_i] - M_i \cdot E[X_j] + M_i \cdot M_j.$$

Here, $E[X_i] = M_i$ and $E[X_j] = M_j$, so $M'_{ij} = M_{ij} - M_i \cdot M_j$. For $i = j$, this is variance V_i, for $i \ne j$, this is covariance C_{ij}. Thus, shift-invariance means that

$$f(\{M_i\}, \{M_{ij}\}) = f(0, \{V_i\}, \{C_{ij}\}).$$

This proves that this combination depends only on the variances and covariances.

Proof of Proposition 5.4. It is easy to check that coefficients of variation and of covariance are scale-invariant, and thus, that any combination of coefficients of variation and covariance is also scale-invariant.

Let us prove that, vice versa, any scale-invariant combination is a function of coefficients of variation and covariance. Indeed, by definition of scale-invariance, the value of this combination should not change if we re-scale the original random vector. In particular, we can re-scale it by dividing each random variable by M_i, i.e., by taking $c_i = 1/M_i$. Then, for the re-scaled random variable $X'_i = X_i/M_i$, the first moments M'_i will be equal to 1. For the second moments, we have

$$M'_{ij} = E[X'_i \cdot X'_j] = E\left[\frac{X_i}{M_i} \cdot \frac{X_j}{M_j}\right] = \frac{E[X_i \cdot X_j]}{M_i \cdot M_j} = \frac{M_{ij}}{M_i \cdot M_j}.$$

For $i = j$, since $M_{ii} = V_i + M_i^2$, we get

$$M'_{ii} = \frac{V_i + M_i^2}{V_i^2} = \frac{V_i}{M_i^2} + 1 = 1 + CV_i^2.$$

For $i \neq j$, since $M_{ij} = C_{ij} + M_i \cdot M_j$, we have

$$M_{ij} = \frac{C_{ij} + M_i \cdot M_j}{M_i \cdot M_j} = \frac{C_{ij}}{M_i \cdot M_j} + 1 = 1 + CV_{ij}.$$

Thus, scale-invariance means that

$$f(\{M_i\}, \{M_{ij}\}) = f(0, \{1 + CV_i^2\}, \{1 + CV_{ij}\}).$$

This proves that this combination depends only on the coefficients of variation and covariance.

Proof of Proposition 5.5. It is easy to check that each correlation is shift- and scale-invariant, and thus, that any function of the correlations is also shift- and scale-invariant.

Let us prove that, vice versa, any shift- and scale-invariant combination is a function of correlations. Indeed, due to Proposition 5.3, since this combination is shift-invariant, it has the form $g(\{V_i\}, \{C_{ij}\})$ for some function g. By definition of scale-invariance, the value of this combination should not change if we re-scale the original random vector. In particular, we can re-scale it by dividing each component X_i by the corresponding standard deviation σ_i, i.e., by taking $c_i = 1/\sigma_i$. After this re-scaling, each difference $X_i - M_i$ is also divided by σ_i. So, for thus re-scaled variables, we have

$$V_i' = E[(X_i' - M_i')^2] = E\left[\frac{X_i - M_i}{\sigma_i} \cdot \frac{X_i - M_i}{\sigma_i}\right] = \frac{E[(X_i - M_i)^2]}{\sigma_i^2} = \frac{V_i}{V_i} = 1$$

and for $i \neq j$, we have

$$C_{ij}' = E[(X_i' - M_i') \cdot (X_j' - M_j')] = E\left[\frac{X_i - M_i}{\sigma_i} \cdot \frac{X_j - M_j}{\sigma_j}\right] =$$

$$\frac{E[(X_i - M_i) \cdot (X_j - M_j)]}{\sigma_i \cdot \sigma_j} = \frac{C_{ij}}{\sigma_i \cdot \sigma_j} = \rho_{ij}.$$

Thus, scale-invariance means that $g(\{V_i\}, \{C_{ij}\}) = g(\{1\}, \{\rho_{ij}\})$.

This proves that this combination depends only on the correlations.

Proof of Proposition 5.6. It is easy to check that skewness is shift- and scale-invariant, and thus, that any function of skewness is also shift- and scale-invariant.

Let us prove that, vice versa, any shift- and scale-invariant combination is a function of skewness. Let us shift X_1 by subtracting M_1 and then re-scale it by dividing the resulting difference $X_1 - M_1$ by σ_1. One can check that for the resulting random variable $X_1' = \dfrac{X_1 - M_1}{\sigma_1}$, we will have $M_1' = 0$, $M_{11}' = 1$, and $M_{111}' = \widetilde{\mu}_3$. Thus, due to shift- and scale-invariance, we have

$$f(M_1, M_{11}, M_{111}) = f(0, 1, \widetilde{\mu}_3).$$

This proves that this combination depends only on the skewness.

Proof of Proposition 5.7 is similar to the proof of Proposition 5.6.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

[2] V. Kreinovich, "Decision making under interval uncertainty (and beyond)", In: P. Guo and W. Pedrycz (eds.), *Human-Centric Decision-Making Models for Social Sciences*, Springer Verlag, 2014, pp. 163–193.

[3] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[4] H. T. Nguyen, O. Kosheleva, and V. Kreinovich, "Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction", *International Journal of Intelligent Systems*, 2009, Vol. 24, No. 1, pp. 27–47.

[5] H. Raiffa, *Decision Analysis*, McGraw-Hill, Columbus, Ohio, 1997.

[6] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.