My Favorite Theorem: Chebyshev's Theorem

Zane Billings
MATH 479: Capstone
19 September, 2019

1 Introduction

Chebyshev's Inequality (or Chebyshev's Theorem, depending on the source) is a useful statement about the spread of random variables, regardless of their distributions. By using the definitions of random variables and their expected value, along with a result called Markov's Inequality (although the result was shown earlier by Chebyshev), Chebyshev's theorem can be proven without the use of measure theory. From applications in the sciences and engineering, to further applications in probability theory (such as the proof of the Weak Law of Large Numbers), Chebyshev's theorem has many uses. The statement of Chebyshev's theorem is as follows: [10]

Theorem 1 (Chebyshev's Theorem)**.** *For any random variable X with mean μ and finite, non-zero variance σ^2,*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

for all strictly positive real numbers k.

Chebyshev's Theorem is my favorite theorem in part due to applications in probability, and also because of its generality. Most introductory statistics courses cover the Empirical Rule, which provides tight bounds for the probability that a given normally distributed random variable takes on a value some distance away from its mean. The Empirical Rule is employed often in the sciences, where many data series are assumed to follow a normal distribution. The beauty of Chebyshev's Theorem lies in the idea that the same (albeit looser) bounds are provided for any random variable, regardless of distribution. The ability to provide bounds for a series of data without knowing anything about the broader distribution is extremely elegant, and I think surprising as well. The generality of Chebyshev's Theorem is what makes the theorem beautiful, and why it is my favorite theorem.

Here, we will provide necessary background and definitions to make the probabilistic statement of Chebyshev's Theorem, present the proofs to useful lemmas and an intermediate result called Markov's Theorem, and finally prove Chebyshev's Theorem. Some applications and the physical interpretation of Chebyshev's Theorem will be discussed.

2 Background

2.1 History

Pafnuty Chebyshev was a Russian mathematican who lived during the nineteenth century. He made contributions to analysis, number theory, and probability theory, and is also known for advising other well-known Russian mathematicians such as Markov and Lyapunov [2].While the development

of probability theory began in the sixteenth century, dating back to Cardano's work [9], a formal axiomatic system of probability theory was not developed until the 1930's by Kolmogorov [8]. Kolmogovorov's axiomatisation is based on measure theory and provides a formal structure for random variables and their probabilities to be expressed. However, Chebyshev's work was published prior to Kolmogorov's axiomatisation [4], and does not require the use of measure theory for a proof. As such, no proofs or definitions are expressed in terms of measure theory.

2.2 Random Variables and their Distributions

Chebyshev's theorem provides a bound for certain probabilities of random variables. Random variables are one of the major fundamental elements of modern probability, and in layman's terms can be (almost tautologically) defined as variables which take on values determined by a random process. While Chebyshev never mentioned "random variables," he still made use of the mathematical construct called a random variable in modern mathematics. As stated previously, an in-depth axiomatic definition of random variables is not needed to understand Chebyshev's work, although one does exist (e.g. as stated by Billingsley in his extremely rigorous work on measure-theoretic probability [1]). For Chebyshev's notion of a random variable we use the following sufficient definition.

Definition 1 (Random variable)**.** A random variable is a function from the set of all possible outcomes of a particular experiment (called the sample space S of an experiment) to the real numbers. [3]

For example, consider the experiment of rolling one six-sided die. The sample space is the set $S = \{1, 2, 3, 4, 5, 6\}$, which are all of the possible values we can get from rolling the die. That is, the set S contains all possible events that can occur as a result of the experiment. We can define a number of random variables which map this simple sample space to the real numbers. If we define X to be 1 if we roll an even number and 0 otherwise, we have defined X as a random variable.

Typically, the capital letter X refers to the random variable itself, while a lowercase x refers to some value which the random variable X can take. Additionally, we say a random variable X is a discrete random variable if the sample space of X contains either a finite number or a countably infinite number of events. If the sample space of X contains an uncountably infinite number of events, X is said to be a continuous random variable. [10] Every such random variable X is equipped with a function called the cumulative distribution function (cdf) of X.

Definition 2 (Cumulative distribution function)**.** The cumulative distribution function, or cdf, of a random variable X, denoted as $F_X(x)$, is defined to be the function

$$F_X(x) = \Pr(X \leq x)$$

for all values x of X, where $\Pr(E)$ denotes the probability of some event E. [3]

While the formal notation is used above, if only one random variable is being considered, the notation for a cdf is conventionally abbreviated to $F(x)$, with the understanding that F is the cdf of X. The cdf of a random variable can be thought of as an "accumulation" function: as x is allowed to increase, the value of the cdf "accumulates" the probability of all of the lower values of x (this description of the cdf is attributed to Dr. Erin McNelis, from Western Carolina University). Due to the definition of the cdf, every cdf will share three major properties. [3]

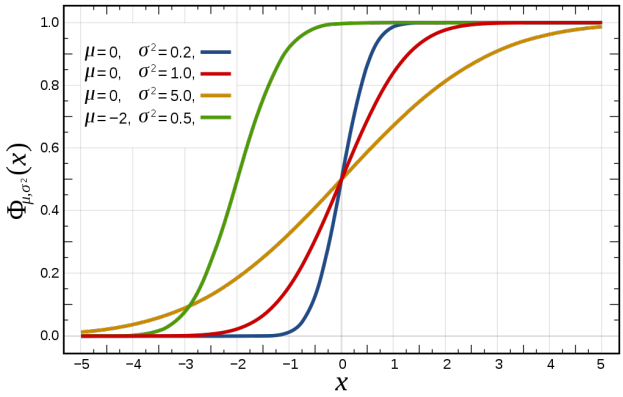

Figure 1: The cumulative distribution function of a normally distributed random variable with varying parameters. The image is in the public domain.

Theorem 2 (Properties of a cdf)**.** *A function F is a cdf if and only if the following properties hold.*

1. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.
2. $F(y) \geq F(x)$ *if* $y > x$; *that is, F is nondecreasing.*
3. $\lim_{x\to x_0^+} F(x) = F(x_0)$ *for all x_0; that is, F is right-continuous everywhere.*

The proof of the above properties depends on the definition of the probability function. [5] The most well-known example is likely the cdf of the normal distribution. While the technical definition of the normal distribution is not necessary for understanding, the cdf of the normal distribution represents the probability that a normal random variable with mean μ and standard deviation σ will attain a value of at least x. The example normal cdfs shown in Figure 1 all display the properties above.

In the case of a continuous random variable, the cdf is a continuous curve like we see in Figure 1. In fact, the cdf provides an alternative definition for continuous and discrete random variables. A random variable can be said to be continuous (more formally, absolutely continuous), if its cdf is continuous everywhere in the real numbers. A random variable is said to be discrete if its cdf is a right-continuous step function on the real numbers. These definitions are equivalent to the previous definitions [3]. A random variable can also have both discrete and continuous pieces, with a piecewise cdf, hence the distinction of an absolutely continuous random variable.

The cdf of a random variable has an intimate relationship with a second distribution function, called the probability mass function (pmf) of the random variable X if X is discrete, or the probability density function (pdf) of X if X is continuous. The pmf of a discrete random variable X is easily stated as $f_X(x) = P(X = x)$ for all values x of X. However, in the continuous case, evaluating the probability of a single point doesn't mean much–on a continuous curve, finding the probability of a single point amounts to integrating the area under the curve with the lower bound equal to the upper bound. In other words, the probability of any one value of a continuous random variable occurring is 0. So, the pdf of a continuous random variable is informally defined as the derivative of the cdf (provided this derivative exists).

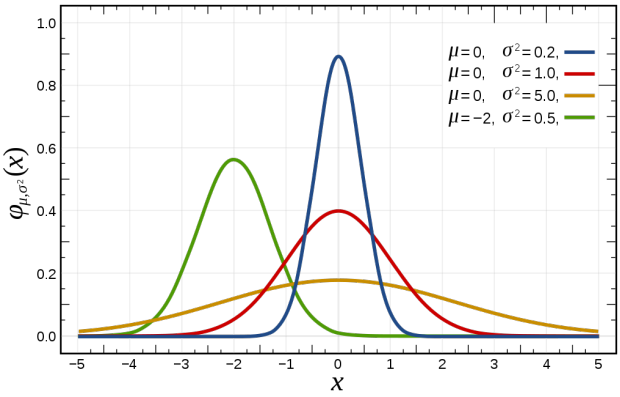

Figure 2: The pdf of the normal distribution for several different parameter values. This function occurs often in many areas of research, and many people are familiar with it. The image is in the public domain.

Definition 3 (Probability density function)**.** The probability density function, or pdf, of a continuous random variable X is the unique function f satisfying

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt$$

for all values x of X. [3]

Note that the probability density function of X is denoted $f_X(x)$, and is commonly abbreviated as simply $f(x)$ if only one random variable has been introduced. The lowercase f is used to indicate the relationship between the pdf and the cdf. While the cdf represents an "accumulation" of probabilities, the pdf represents the probability that a random variable takes on a specific range of values–while the probability that, say, a random variable X takes on the value a is 0, the probability that $a < X < b$ is nonzero and can be represented by the pdf. Similarly to the cdf, the pdf of a random variable is defined by certain properties.

Theorem 3 (Properties of a pdf)**.** *A function $f_X(x)$ is a pdf if and only if these properties hold.*

1. $f_X(x) \geq 0$ *for all values x of X.*
2. $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$. *[3]*

The properties can be proven using the definition of the pdf. [3] The most common example for a pdf is also the function of the normal distribution. Most scientists and mathematicians are familiar with the bell curve produced by a normal distribution: this bell curve is actually the pdf of the normal distribution. Figure 2 shows the normal pdf using several different sets of parameter values. (As a side note, the normal pdf cannot be integrated using elementary functions–the normal cdf is actually defined in terms of the error function.)

Additionally, note that every random variable has a support set whose definition relies on the distribution function (or the mass function in the discrete case).

Definition 4 (Support set)**.** The support set, $\mathcal{S}$, (or simply the support) of a random variable X, is the subset of the sample space of X (i.e. the domain of X) containing all values of the domain for which the value of the pdf of X is strictly positive. [3]

Importantly, since every random variable has a support, we can say for a random variable X,

$$\int_{\mathbb{R}} f_X(x)\ dx = \int_{-\infty}^{\infty} f_X(x)\ dx = \int_{\mathcal{S}} f_X(x)\ dx = 1.$$

From here on, we will only discuss absolutely continuous random variables: in most cases, the proof of a result for a discrete random variable immediately follows for a proof utilizing continuous random variables if either the discrete random variable is thought of as a piecewise function where each piece is a constant function, or if all definitions involving continuous random variables are replaced with the equivalent definitions for discrete random variables. While the proofs in the discrete case are not trivial, they typically follow the same structure.

2.3 Expectation and Variance

The expected value of a random variable is a generalization of the mean which is applicable to random variables. The expected variable represents the long-term running average of the values of the random variable. In fact, the Law of Large Numbers states that the running average of the values of a random variable should approach the expected value as the number of trials increases to infinity [10].

Definition 5 (Expected value of a continuous random variable)**.** The expected value, or expectance, of a continuous random variable X, denoted $\mathbb{E}[X]$ or μ_X is defined to be

$$\mu_X = \mathbb{E}[X] = \int_{\mathcal{S}} x f_X(x)\ dx. \text{ [10]}$$

While the expected value encodes information about the average value of a random variable, the variance of a random variable provides information about how far away the value a random variable takes on tends to be from the expected value. In other words, the variance provides information about the spread of a random variable while the expected value provides information about the center of a random variable. The variance of a random variable is defined as the expected value of the squared difference between the value of a random value and the expected value.

Definition 6. The variance of a continuous random variable X, denoted σ_X^2 or $\text{Var}[X]$ is defined as

$$\sigma_X^2 = \text{Var}[X] = \mathbb{E}[(X-\mu)^2] = \int_{\mathcal{S}} (x-\mu)^2 f_X(x)\ dx. \text{ [10]}$$

These quantities are visualized in Figure 2: the mean of the normal pdf is where the center of the bell curve is located, while the standard deviation tells how far the curve spreads out. Note that the standard deviation of a random variable is denoted as σ_X and is defined to be the square root of the variance.

2.4 Jointly Distributed Random Variables

So far, we have only considered the case where just one random variable is considered in isolation. However, for a later result we will need to consider two random variables at the same time, and many real world applications require several random variables to be considered simultaneously. When we are interested in the values that two different random variables, say X and Y, take on at the same define, we can define the joint distribution of X and Y.

Definition 7 (Joint distribution function)**.** The joint distribution function of two continuous random variables X and Y is the multivariate function F defined as

$$F(x, y) = \Pr(X \leq x, Y \leq b),$$

where x is a value of X and y is a value of Y. [7] Note that $\Pr(e_1, e_2)$ denotes the probability that events e_1 and e_2 both occur.

In the case of two jointly distributed continuous random variables, X and Y, the joint distribution is equipped with a joint probability density function as well, which gives the probability that X takes on some value within a specified interval and Y takes on a value within a specified interval at the same time.

Definition 8 (Joint probability density function)**.** The joint probability density function of two absolutely continuous random variables X and Y is defined as the unique non-negative function f which satisfies the relationship

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\ dy\ dx$$

for all values x of X and y of Y. When this function f exists, X and Y are called jointly continuous random variables. [10]

More importantly for the purpose of proving Chebyshev's theorem, if two random variables are jointly distributed, information about each of those random variables can be extracted from the joint cdf in the form of the marginal density functions.

Definition 9 (Marginal density function)**.** Let X and Y be jointly continuous random variables which have the joint density function f. Then, the marginal density function of X is

$$f_X(x) = \int_{\mathcal{S}_Y} f(x, y)\ dy,$$

which is a function of X only (the notation $\mathcal{S}_Y$ refers to the support set of the random variable Y). Similarly, the marginal density function of Y is

$$f_Y(y) = \int_{\mathcal{S}_X} f(x, y)\ dx,$$

which is likewise a function of Y only. [10, 7]

The marginal density functions of two jointly continuous random variables will be equivalent to the univariate pdf of each of the random variables. [3] The expected value of two jointly continuous random variables can be computed using the marginal density functions.

Theorem 4 (Linearity of Expected Value)**.** *Let X and Y be jointly continuous random variables, and let a and b be real numbers. Then,*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

This property is called the linearity of expected value (that is, expected value is a linear operator).

Proof. Let X and Y be jointly continuous random variables with the joint density function f and respective marginal density functions f_X and f_Y, and let a and b be arbitrary real numbers. By definition, we have that

$$\mathbb{E}[aX + bY] = \int_{\mathcal{S}_X} \int_{S_Y} (ax + by) f(x, y) \; dy \; dx.$$

Since integration is a linear operator, we can break integrals apart at sums, so we have that

$$\mathbb{E}[aX + bY] = \int_{\mathcal{S}_X} \int_{S_Y} ayf(x, y) \; dy \; dx + \int_{\mathcal{S}_X} \int_{S_Y} byf(x, y) \; dy \; dx.$$

Then, constant multiples can be "pulled out" of integrals by the same property (the linearity of integration), giving

$$\mathbb{E}[aX + bY] = a \int_{\mathcal{S}_X} \int_{S_Y} xf(x, y) \; dy \; dx + b \int_{\mathcal{S}_X} \int_{S_Y} yf(x, y) \; dy \; dx.$$

Since both of these integrals converge to finite numbers, Fubini's theorem permits us to change the order of integration of the second integral [6], and we have

$$\mathbb{E}[aX + bY] = a \int_{\mathcal{S}_X} \int_{S_Y} xf(x, y) \; dy \; dx + b \int_{S_Y} \int_{\mathcal{S}_X} yf(x, y) \; dx \; dy.$$

Now, since we are integrating with respect to y in the first integral, x can be treated as a constant. Likewise in the second integral, we are integrating with respect to x, so y can be treated as a constant. So again by the linearity of integration, we have

$$\mathbb{E}[aX + bY] = a \int_{\mathcal{S}_X} x \int_{S_Y} f(x, y) \; dy \; dx + b \int_{S_Y} y \int_{\mathcal{S}_X} f(x, y) \; dx \; dy.$$

Applying the definition of the marginal distribution function, we get

$$\mathbb{E}[aX + bY] = a \int_{\mathcal{S}_X} xf_X(x, y) \; dx + b \int_{S_Y} yf_Y(x, y) \; dy,$$

and thus we have, by the definition of expected value,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

This is exactly what we wanted to show, concluding the proof. □

With the expected value of jointly continuous random variables proven, we are now equipped with all of the essential tools we need to prove Chebyshev's Theorem. The proof of Chebyshev's Theorem will require the proof of two lemmas and an intermediate result, Markov's Inequality.

3 Intermediate Results

Given all of the previous definitions and results, we now present three intermediate results. The first lemma is used to prove the second, and the second lemma will be used to prove Markov's Inequality, a strong result which is used in the proof of Chebyshev's Theorem. While Chebyshev's Theorem can be proven directly, the proof using Markov's Inequality is extremely similar, and provides a framework for other results in probability theory.

Lemma 1. *Let X be a continuous random variable, and let a be a real number. If* $\Pr(X \geq a) = 1$*, then* $\mathbb{E}[X] \geq a$*.*

Proof. Suppose X is a continuous random variable and a is a real number such that $\Pr(X \geq a) = 1$. Furthermore, let $\mathcal{S}$ be the support set of X. Since $\Pr(X \geq a) = 1$ we will show that $x > a$ for every value x of X.

Now, suppose that we have some value x_i of X such that $x_i < a$ and $x_i \in \mathcal{S}$. But, since $\Pr(X \geq a) = 1$, we have by necessity that $\Pr(X < a) = 0$, and since $x_i < a$, we have that $\Pr(X < x_i) = 0$, since $F_X(x)$ is nondecreasing (2). However, this contradicts the assumption that $x_i \in \mathcal{S}$, so we see that there cannot be a value x of X where both $x_i < a$ and $x_i \in \mathcal{S}$ are true. Thus, $x \geq a$ for every $x \in \mathcal{S}$.

With this established, consider the expected value of X. Since $x \geq a$, we have that

$$\mathbb{E}[X] = \int_{\mathcal{S}} x f_X(x)\ dx \geq \int_{\mathcal{S}} a f_X(x)\ dx.$$

Since integration is linear, we have

$$\mathbb{E}[X] \geq a \int_{\mathcal{S}} f_X(x)\ dx,$$

and thus, by Theorem 3,

$$\mathbb{E}[X] \geq a,$$

which is what we desired. □

Lemma 2. *Let X and Y be two continuous random variables. If* $\Pr(X \geq Y) = 1$*, then* $\mathbb{E}[X] \geq \mathbb{E}[Y]$*.*

Proof. Let X and Y be two continuous random variables such that $\Pr(X \geq Y) = 1$. Additionally, let Z be a continuous random variable such that $Z = X - Y$.

By rearranging the given probability statement, we have that $\Pr(X - Y \geq 0) = 1$, and thus that $\Pr(Z \geq 0) = 1$. Applying Lemma 1, we see that $\mathbb{E}[Z] \geq 0$, and thus that $E[X - Y] \geq 0$. By Theorem 4 (the linearity of expected value), we see that $E[X - Y] = E[X] - E[Y] \geq 0$, and thus that $E[X] \geq E[Y]$. □

Theorem 5 (Markov's Inequality)**.** *For a non-negative continuous random variable X and strictly positive real number a,*

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Suppose X is a non-negative random variable (that is, a random variable which only takes on values that are nonnegative) and a is a strictly positive real number such that $\Pr(X \geq a)$.

Now, we construct the indicator variable I_a for the event $X \geq a$. That is,

$$I_a = \begin{cases} 1, \text{ if } X \geq a; \\ 0, \text{ if } X < a. \end{cases}$$

Note that by construction,

$$\mathbb{E}[I_a] = 1(\Pr(X \geq a)) + 0(\Pr(X < a)) = \Pr(X \geq a).$$

Next, we will demonstrate that the quantity $aI_a \leq x$ for every value x of X. We have two cases where this needs to be shown: the cases where $X \geq a$, and the case where $X < a$.

Consider first the case when $X \geq a$. Then, $I_a = 1$ by definition. So, $aI_a = a$, and hence $X \geq aI_a$. Now consider when $X \leq a$. Then, $I_a = 0$ by definition, and thus $aI_a = 0$. However, X is nonnegative by assumption, so we have $X \geq 0 = aI_a$. So, we see that $X \geq aI_a$ in every case, and thus $\Pr((X \geq aI_a) = 1$.

Then, letting $Y = aI_a$, we have $\Pr(X \geq Y) = 1$, and by Lemma 2,

$$\mathbb{E}[X] \geq \mathbb{E}[Y] = \mathbb{E}[aI_a].$$

Applying Theorem 4 (Linearity), we have

$$\mathbb{E}[X] \geq a\mathbb{E}[I_a] = a\Pr(X \geq a).$$

Rearranging, we get

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

This is what we wanted to show, so this concludes the proof. □

Notably, Markov's Inequality does not itself require X to be a continuous random variable. However, since all previous results and definitions were stated in the continuous case, here Markov's Inequality is only stated to be true for the continuous (and specifically, the absolutely continuous) case.

4 Proof of Chebyshev's Theorem

With all three of these intermediate results stated and proven, we finally have Markov's Inequality, the major tool we need to prove Chebyshev's Theorem. For the proof, we first show that we can make a substitution satisfying the requirements of Markov's Theory, and then simplify the resulting inequality to obtain the statement of Chebyshev's Theorem.

Theorem 6. *For any random variable X with finite mean μ and finite, non-zero variance σ^2,*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

for all strictly positive real numbers k.

Proof. Let X be a continuous random variable with finite mean μ and finite, non-zero variance σ^2, and let $Y = (X - \mu)^2$. Finally, let k be a non-zero real number.

Note that Y is a nonnegative random variable, regardless of the value of X, and k^2 is strictly positive. Since k^2 and σ^2 will both be strictly positive, $k^2\sigma^2$ is also strictly positive.

Then, by Markov's Inequality, we have that

$$\Pr(Y \geq \sigma^2 k^2) \leq \frac{\mathbb{E}[Y]}{\sigma^2 k^2}.$$

Since $Y = (X - \mu)^2$, we have that $\mathbb{E}[Y] = \mathrm{Var}[X]$ by the definition of variance. Thus,

$$\Pr((X - \mu)^2 \geq \sigma^2 k^2) \leq \frac{\mathrm{Var}[X]}{\sigma^2 k^2}.$$

Notably, $\text{Var}[X] = \sigma^2$. Making this substitution and simplifying, we have

$$\Pr(|X - \mu| \geq \sigma k) \leq \frac{1}{k^2}.$$

This is the statement of Chebyshev's Theorem, so we are done. □

In simple language, Chebyshev's Inequality says that for any random variable, the probability that the variable takes on a value more than k standard deviations from its mean is no greater than $1/k^2$, and thus bounds the proportion of the distribution's values that must lie within k standard deviations of the mean regardless of what the distribution is.

5 Conclusion

Chebyshev's inequality provides weaker bounds than, say, the Empirical Rule for the normal distribution. For a normally distributed random variable X, Chebyshev's theorem guarantees that at least 75% of the values of the distribution will be within two standard deviations of the mean, while the Empirical Rule guarantees that about 95% of the values of the distribution will lie within two standard deviations of the mean.

In this respect, Chebyshev's Theorem is a fairly weak bound. But, the beauty and the usefulness of Chebyshev's Inequality comes from the fact that Chebyshev's Theorem is true for any random variable, regardless of distribution, so long as the random variable has a mean and a finite variance. The generality of Chebyshev's Theorem makes it a beautiful and useful statement.

References

[1] Patrick Billingsley. *Probability and Measure.* John Wiley & Sons, third edition, 1995.

[2] Paul Butzer and François Jongmans. P. L. Chebyshev (1821–1894): A guide to his life and work. *Journal of Approximation Theory*, 96, 1998.

[3] George Casella and Roger L. Berger. *Statistical Inference.* Duxbury, second edition, 2002.

[4] P. Chebyshev. Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, 2, 1867.

[5] Robert V. Hogg and Allen T. Craig. *Introduction to Mathematical Statistics.* Macmillan Publishing Co., Inc., fourth edition, 1978.

[6] L. D. Kudryavtsev. Fubini theorem. In Michiel Hazewinkel, editor, *Encyclopedia of Mathematics.* Springer, 2001.

[7] Sheldon M. Ross. *Introduction to probability models.* Academic Press, Inc., fourth edition, 1989.

[8] Glenn Shafer and Vladimir Vovk. The sources of Kolmogorov's Grundbegriffe. *Statistical Science*, 26, 2006.

[9] I. Todhunter. *A history of the mathematical theory of probability.* Macmillian and Co., 1865.

[10] Dennis D. Wackerly, William Medenhall III, and Richard L. Scheaffer. *Mathematical Statistics with Applications.* Thomson, seventh edition, 2008.