# 2. Variance and Higher Moments

Recall that by taking the expected value of various transformations of a random variable, we can measure many interesting characteristics of the distribution of the variable. In this section, we will study expected values that measure spread, skewness and other properties.

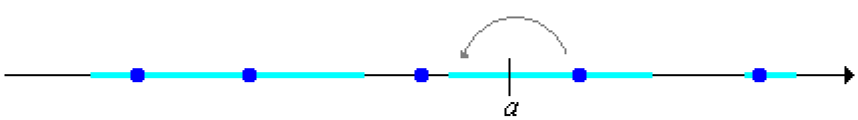
## Variance

### Definitions

As usual, we start with a random experiment with probability measure $\mathbb{P}$ on an underlying sample space. Suppose that $X$ is a random variable for the experiment, taking values in $S \subseteq \mathbb{R}$. Recall that the expected value or mean of $X$ gives the center of the distribution of $X$. The **variance** of $X$ is a measure of the spread of the distribution about the mean and is defined by

$$\operatorname{var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

Recall that the **second moment of $X$ about $a$** is $\mathbb{E}\left((X - a)^2\right)$. Thus, the variance is the second moment of $X$ about $\mu = \mathbb{E}(X)$, or equivalently, the **second central moment** of $X$. Second moments have a nice interpretation in physics, if we think of the distribution of $X$ as a mass distribution in $\mathbb{R}$. Then the second moment of $X$ about $a$ is the **moment of inertia** of the mass distribution about $a$. This is a measure of the resistance of the mass distribution to any change in its rotational motion about $a$. In particular, the variance of $X$ is the moment of inertia of the mass distribution about the center of mass $\mu$.




1. Suppose that $X$has a discrete distribution with probability density function $f$. Use the change of variables theorem to show that

$$\operatorname{var}(X) = \sum_{x \in S} (x - \mathbb{E}(X))^2 f(x)$$

2. Suppose that $X$ has a continuous distribution with probability density function $f$. Use the change of variables theorem to show that

$$\operatorname{var}(X) = \int_S (x - \mathbb{E}(X))^2 f(x) dx$$

The **standard deviation** of $X$ is the square root of the variance. It also measures dispersion about the mean but has the same physical units as the variable $X$.

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$

### Properties

The following exercises give some basic properties of variance, which in turn rely on basic properties of expected value:

3. Show that $\mathrm{var}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$.

4. Show that $\mathrm{var}(X) \ge 0$

5. Show that $\mathrm{var}(X) = 0$ if and only if $\mathbb{P}(X = c) = 1$ for some constant $c$.

6. Show that if $a$ and $b$ are constants then $\mathrm{var}(a\,X + b) = a^2\,\mathrm{var}(X)$.

7. Show that the random variable given below has mean 0 and variance 1:

$$Z = \frac{X - \mathbb{E}(X)}{\mathrm{sd}(X)}$$

.

The random variable $Z$ in Exercise 7 is sometimes called the **standard score** associated with $X$. Since $X$ and its mean and standard deviation all have the same physical units, the standard score $Z$ is dimensionless. It measures the directed distance from $\mathbb{E}(X)$ to $X$ in terms of standard deviations.

On the other hand, when $X \ge 0$, the ratio of standard deviation to mean is called the **coefficient of variation**. This quantity also is dimensionless, and is sometimes used to compare variability for random variables with different means.

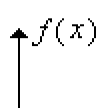
$$\mathrm{cv}(X) = \frac{\mathrm{sd}(X)}{\mathbb{E}(X)}$$

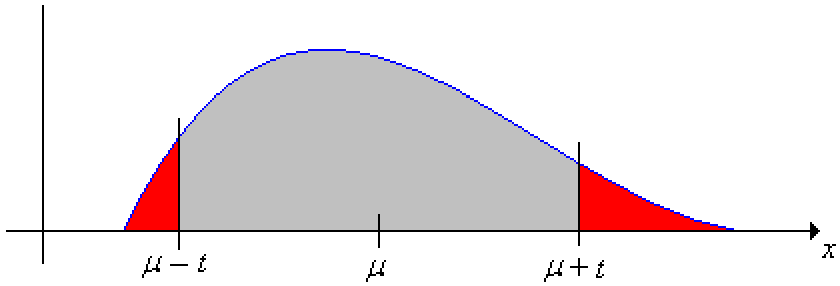
### Chebyshev's Inequality

**Chebyshev's inequality** (named after Pafnuty Chebyshev) gives an upper bound on the probability that a random variable will be more than a specified distance from its mean. This is often useful in applied problems where the distribution is unknown, but the mean and variance are at least approximately known. In the following two exercises, suppose that $X$ is a real-valued random variable with mean $\mu = \mathbb{E}(X)$ and standard deviation $\sigma = \mathrm{sd}(X)$.

8. Use Markov's inequality to prove Chebyshev's inequality:

$$\mathbb{P}(|X - \mu| \ge t) \le \frac{\sigma^2}{t^2} \text{ for } t > 0$$






9. Establish the following equivalent version of Chebyshev's inequality:

$$\mathbb{P}(|X - \mu| \ge k\,\sigma) \le \frac{1}{k^2} \text{ for } k > 0$$

The usefulness of the Chebyshev inequality comes from the fact that it holds for any distribution (assuming only that the mean and variance exist). The tradeoff is that for many specific distributions, the Chebyshev bound is rather crude. Note in particular that in the last exercise, the bound is useless when $k \le 1$, since 1 is an upper bound for the probability of any event.

## Examples and Applications

### Indicator Variables

10. Suppose that $X$ is an indicator variable with $\mathbb{P}(X = 1) = p$.

a. Recall that $\mathbb{E}(X) = p$.
b. Show that $\text{var}(X) = p\,(1 - p)$.
c. Sketch the graph of $\text{var}(X)$ as a function of $p$.

Note that the minimum value of $\text{var}(X)$ is 0, and occurs when $p = 0$ and $p = 1$. The maximum value is $\frac{1}{4}$ and occurs when $p = \frac{1}{2}$

### Uniform Distributions

11. Suppose that $X$ has the discrete uniform distribution on the integer interval $\{m, m + 1, ..., n\}$ where $m \le n$.

a. Recall that $\mathbb{E}(X) = \frac{1}{2}\,(m + n)$.
b. Show that $\text{var}(X) = \frac{1}{12}\,(n - m)\,(n - m + 2)$.

12. Suppose that $X$ has the continuous uniform distribution on the interval $[a, b]$.

a. Recall that $\mathbb{E}(X) = \frac{1}{2}\,(a + b)$.

b. Show that $\text{var}(X) = \frac{1}{12}(b - a)^2$.

Note that in both the discrete and continuous cases, the variance depends only on the length of the interval.

### Dice

Recall that a **standard die** is a six-sided die. A **fair** die is one in which the faces are equally likely. An **ace-six flat die** is a standard die in which faces 1 and 6 have probability $\frac{1}{4}$ each, and faces 2, 3, 4, and 5 have probability $\frac{1}{8}$ each.

13. A standard, fair die is thrown. Find the mean, variance, and standard deviation of the score.

14. In the dice experiment, select one fair die. Run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

15. An ace-six flat die is thrown. Find the mean, variance and standard deviation.

16. In the dice experiment, select one ace-six flat die. Run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

### The Poisson Distribution

Recall that the **Poisson distribution** has density function

$$f(n) = e^{-a}\frac{a^n}{n!}, \quad n \in \mathbb{N}$$

where $a > 0$ is a parameter. The Poisson distribution is named after Simeon Poisson and is widely used to model the number of "random points" in a region of time or space; the parameter $a$ is proportional to the size of the region. The Poisson distribution is studied in detail in the chapter on the Poisson Process.

17. Suppose that $N$ has the Poisson distribution with parameter $a$.

a. Recall that $\mathbb{E}(N) = a$.
b. Show that $\text{var}(N) = a$.

Thus, the parameter is both the mean and the variance of the distribution.

18. In the Poisson experiment, the parameter is $a = r\,t$. Vary the parameter and note the size and location

of the mean-standard deviation bar. For selected values of the parameter, run the experiment 1000 times updating every 10 runs. Note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

### The Geometric Distribution

Recall that the **geometric distribution** on $\mathbb{N}_+$ is a discrete distribution with density function

$$f(n) = p\,(1 - p)^{n-1}, \;\; n \in \mathbb{N}_+$$

where $p \in (0, 1]$ is a parameter. The geometric distribution governs the trial number of the first success in a sequence of Bernoulli trials with success parameter $p$.

19. Suppose that $W$ has the geometric distribution with success parameter $p$.

a. Recall that $\mathbb{E}(W) = \frac{1}{p}$.
b. Show that $\text{var}(W) = \frac{1-p}{p^2}$.

20. In the negative binomial experiment, set $k = 1$ to get the geometric distribution . Vary $p$ with the scroll bar and note the size and location of the mean-standard deviation bar. For selected values of $p$, run the experiment 1000 times updating every 10 runs. Note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

21. Suppose that $W$ has the geometric distribution with parameter $p = \frac{3}{4}$. Compute the true value and the Chebyshev bound for the probability that $Y$ is at least 2 standard deviations away from the mean.

### The Exponential Distribution

Recall that the **exponential distribution** is a continuous distribution with probability density function

$$f(t) = r\,e^{-r\,t}, \;\; t \ge 0$$

where $r > 0$ is the with **rate parameter**. This distribution is widely used to model failure times and other “arrival times”. The exponential distribution is studied in detail in the chapter on the Poisson Process.

22. Suppose that $X$ has the exponential distribution with rate parameter $r$.

a. Recall that $\mathbb{E}(X) = \frac{1}{r}$.
b. Show that $\text{sd}(X) = \frac{1}{r}$.

Thus, for the exponential distribution, the mean and standard deviation are the same.

23. In the gamma experiment, set $k = 1$ to get the exponential distribution. Vary $r$ with the scroll bar and note the size and location of the mean-standard deviation bar. For selected values of $r$, run the experiment 1000 times updating every 10 runs. Note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

24. Suppose that $X$ has the exponential distribution with rate parameter $r > 0$. Compute the true value and the Chebyshev bound for the probability that $X$ is at least $k$ standard deviations away from the mean.

### The Pareto Distribution

Recall that the **Pareto distribution** is a continuous distribution with density function

$$f(x) = \frac{a}{x^{a+1}}, \quad x \geq 1$$

where $a > 0$ is a parameter. The Pareto distribution is named for Vilfredo Pareto. It is a heavy-tailed distribution that is widely used to model financial variables such as income. The Pareto distribution is studied in detail in the chapter on Special Distributions.

25. Suppose that $X$ has the Pareto distribution with shape parameter $a$.

a. Recall that $\mathbb{E}(X) = \begin{cases} \infty, & a \in (0, 1] \\ \frac{a}{a-1}, & a \in (1, \infty) \end{cases}$

b. Show that $\text{var}(X) = \begin{cases} \text{undefined}, & a \in (0, 1] \\ \infty, & a \in (1, 2] \\ \frac{a}{(a-1)^2 (a-2)}, & a \in (2, \infty] \end{cases}$

26. In the random variable experiment, select the Pareto distribution. Vary $a$ with the scroll bar and note the size and location of the mean/standard deviation bar. For each of the following values of $a$, run the experiment 1000 times updating every 10 runs and note the behavior of the empirical mean and standard deviation.

a. $a = 1$
b. $a = 2$
c. $a = 3$

### The Normal Distribution

Recall that the **standard normal distribution** is a continuous distribution with density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad z \in \mathbb{R}$$

Normal distributions are widely used to model physical measurements subject to small, random errors and are studied in detail in the chapter on Special Distributions.

27. Suppose that $Z$ has the standard normal distribution.

a. Recall that $\mathbb{E}(Z) = 0$.
b. Show that $\text{var}(Z) = 1$. *Hint*: Integrate by parts in the integral for $\mathbb{E}(Z^2)$.

28. Suppose again that $Z$ has the standard normal distribution and that $\mu \in (-\infty, \infty)$, $\sigma \in (0, \infty)$. Recall that $X = \mu + \sigma Z$ has the normal distribution with **location parameter** $\mu$ and **scale parameter** $\sigma$.

a. Recall that $\mathbb{E}(X) = \mu$
b. Show that $\text{var}(X) = \sigma^2$

Thus, as the notation suggests, the location parameter $\mu$ is also the mean and the scale parameter $\sigma$ is also the standard deviation.

29. In the random variable experiment, select the normal distribution. Vary the parameters and note the shape and location of the mean-standard deviation bar. For selected parameter values, run the experiment 1000 times updating every 10 runs and note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

### Beta Distributions

The distributions in this subsection belong to the family of **beta distributions**, which are widely used to model random proportions and probabilities. The beta distribution is studied in detail in the chapter on Special Distributions.

30. Graph the density functions below and compute the mean and variance of each.

a. $f(x) = 6x(1 - x), \ 0 \le x \le 1$
b. $f(x) = 12x^2(1 - x), \ 0 \le x \le 1$
c. $f(x) = 12x(1 - x)^2, \ 0 \le x \le 1$
d. $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \ 0 < x < 1$.

The particular beta distribution in part (d) is also known as the **arcsine distribution**.

### Miscellaneous Exercises

31. Suppose that $X$ is a real-valued random variable with $\mathbb{E}(X) = 5$ and $\text{var}(X) = 4$. Find each of the following:

a. $\text{var}(3\,X - 2)$
b. $\mathbb{E}\left(X^2\right)$

32. Suppose that $X_1$ and $X_2$ are independent, real-valued random variables with $\mathbb{E}(X_i) = \mu_i$ and $\text{var}(X_i) = \sigma_i$ for $i \in \{1, 2\}$. Show that

$$\text{var}(X_1\,X_2) = \left({\sigma_1}^2 + {\mu_1}^2\right)\left({\sigma_2}^2 + {\mu_2}^2\right) - {\mu_1}^2\,{\mu_2}^2$$

33. Marilyn Vos Savant has an IQ of 228. Assuming that the distribution of IQ scores has mean 100 and standard deviation 15, find Marilyn's standard score.

## Higher Moments

Suppose that $X$ is a real-valued random variable. Recall again that the variance of $X$is the second moment of $X$ about the mean, and measures the spread of the distribution of $X$ about the mean. The third and fourth moments of $X$ about the mean also measure interesting features of the distribution. The third moment measures *skewness*, the lack of symmetry, while the fourth moment measures *kurtosis*, the degree to which the distribution is peaked. The actual numerical measures of these characteristics are standardized to eliminate the physical units, by dividing by an appropriate power of the standard deviation. As usual, we assume that all expected values given below exist, and we will let $\mu = \mathbb{E}(X)$ and $\sigma = \text{sd}(X)$.

### Skewness

The **skewness** of $X$ is defined to be

$$\text{skew}(X) = \frac{\mathbb{E}\left((X - \mu)^3\right)}{\sigma^3}$$

34. Suppose that $X$ has a continuous distribution with probability density $f$ that is symmetric about $a$: $f(a + t) = f(a - t)$, $t \in \mathbb{R}$.

a. Recall that $\mathbb{E}(X) = a$.
b. Show that $\text{skew}(X) = 0$.

35. Show that

$$\text{skew}(X) = \frac{\mathbb{E}\left(X^3\right) - 3\,\mu\,\mathbb{E}\left(X^2\right) + 2\,\mu^3}{\sigma^3}$$

### Kurtosis

The **kurtosis** of $X$ is defined to be

$$\text{kurt}(X) = \frac{\mathbb{E}\left((X - \mu)^4\right)}{\sigma^4}$$

36. Show that

$$\text{kurt}(X) = \frac{\mathbb{E}\left(X^4\right) - 4\,\mu\,\mathbb{E}\left(X^3\right) + 6\,\mu^2\,\mathbb{E}\left(X^2\right) - 3\,\mu^4}{\sigma^4}$$

### Exercises

37. Suppose that $X$ has uniform distribution on the interval $[a, b]$. Compute each of the following:

a. $\text{skew}(X)$
b. $\text{kurt}(X)$

38. Suppose that $X$ has the exponential distribution with rate parameter $r > 0$. Compute each of the following:

a. $\text{skew}(X)$
b. $\text{kurt}(X)$

39. Suppose that $X$ has the Pareto distribution with shape parameter $a > 4$. Compute each of the following:

a. $\text{skew}(X)$
b. $\text{kurt}(X)$

40. Suppose that $Z$ has the standard normal distribution. Compute each of the following:

a. $\text{skew}(Z)$
b. $\text{kurt}(Z)$

41. Graph the following density functions and compute the mean, variance, skewness and kurtosis of each. (The corresponding distributions are all members of the family of beta distributions).

a. $f(x) = 6\,x\,(1 - x),\ \ 0 \le x \le 1$

b. $f(x) = 12\,x^2\,(1 - x),\ 0 \le x \le 1$
c. $f(x) = 12\,x\,(1 - x)^2,\ 0 \le x \le 1$
d. $f(x) = \frac{1}{\pi\,\sqrt{x\,(1 - x)}}$, $0 < x < 1$, the arcsine distribution.

## Vector Space Concepts

Variance and higher moments are related to the concept of norm and distance in the theory of vector spaces. This connection can help unify and illuminate some of the ideas.

Our **vector space** $\mathcal{V}$ consists of all real-valued random variables defined on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (that is, relative to a given random experiment). Recall that two random variables are equivalent if they are equal with probability 1. We consider two such random variables as the same vector, so that technically, our vector space consists of **equivalence classes** under this equivalence relation. The **addition operator** corresponds to the usual addition of two real-valued random variables, and the operation of **scalar multiplication** corresponds to the usual multiplication of a real-valued random variable by a real (non-random) number.

### Norm

Let $X$ be a real-valued random variable. For $k \ge 1$, we define the ***k*-norm** by

$$\|X\|_k = \mathbb{E}\left(|X|^k\right)^{1/k}$$

Thus, $\|X\|_k$ is a measure of the size of $X$ in a certain sense. The following exercises establish the fundamental properties.

42. Show that $\|X\|_k \ge 0$ for any $X$.

43. Show that $\|X\|_k = 0$ if and only if $\mathbb{P}(X = 0) = 1$ (so that $X$ is equivalent to 0).

44. Show that $\|c\,X\|_k = |c|\,\|X\|_k$ for any constant $c$.

The next exercise gives **Minkowski's inequality**, named for Hermann Minkowski. It is also known as the **triangle inequality**.

45. Show that $\|X + Y\|_k \le \|X\|_k + \|Y\|_k$ for any $X$ and $Y$.

a. Show that $S = \left\{(x, y) \in \mathbb{R}^2 : (x \ge 0) \text{ and } (y \ge 0)\right\}$ is a convex set and that $g(x, y) = \left(x^{1/k} + y^{1/k}\right)^k$ is concave on $S$.
b. Use (a) and Jensen's inequality to conclude that if $U$ and $V$ are nonnegative random variables then

$$\mathbb{E}\left(\left(U^{1/k} + V^{1/k}\right)^k\right) \le \left(\mathbb{E}(U)^{1/k} + \mathbb{E}(V)^{1/k}\right)^k$$

c. In (b) let $U = |X|^k$ and $V = |Y|^k$ and then do some algebra.

It follows from Exercises 42-45 that the set of random variables with finite $k^{\text{th}}$ moment forms a **subspace** of our parent vector space $\mathcal{V}$, and that the $k$-norm really is a norm on this vector space:

$$\mathcal{V}_k = \{X \in \mathcal{V} : \|X\|_k < \infty\}$$

Our next exercise gives **Lyapunov's inequality**, named for Aleksandr Lyapunov. This inequality shows that the $k$-norm of a random variable is increasing in $k$.

46. Show that if $j \le k$ then $\|X\|_j \le \|X\|_k$.

a. Show that $S = \{x \in \mathbb{R} : x \ge 0\}$ is convex, and $g(x) = x^{k/j}$ is convex on $S$.
b. Use part (a) and Jensen's inequality to conclude that if $U$ is a nonnegative random variable then $\mathbb{E}(U)^{k/j} \le \mathbb{E}\left(U^{k/j}\right)$.
c. In (b), let $U = |X|^j$ and do some algebra.

Lyapunov's inequality shows that if $j \le k$ and $X$ has a finite $k^{\text{th}}$ moment, then $X$ has a finite $j^{\text{th}}$ moment as well. Thus, $\mathcal{V}_k$ is a subspace of $\mathcal{V}_j$.

47. Suppose that $X$ is uniformly distributed on the interval $[0, 1]$.

a. Find $\|X\|_k$.
b. Graph $\|X\|_k$ as a function of $k$.
c. Find $\lim_{k \to \infty} \|X\|_k$.

48. Suppose that $X$ has probability density function $f(x) = \frac{a}{x^{a+1}}, \ x \ge 1$, where $a > 0$ is a parameter. Thus, $X$ has the Pareto distribution with shape parameter $a$.

a. Find $\|X\|_k$.
b. Graph $\|X\|_k$ as a function of $k < a$.
c. Find $\lim_{k \uparrow a} \|X\|_k$.

49. Suppose that $(X, Y)$ has probability density function $f(x, y) = x + y, \ 0 \le x \le 1, 0 \le y \le 1$. Verify Minkowski's inequality.

### Distance

The $k$-norm, like any norm on a vector space, can be used to measure distance; we simply compute the norm of the difference between two vectors. Thus, we define the ***k*-distance** (or ***k*-metric**) between real-valued random variables $X$ and $Y$ to be

$$d_k(X, Y) = \|Y - X\|_k = \mathbb{E}\left(|Y - X|^k\right)^{1/k}$$

.

The properties in the following exercises are analogies of the properties in Exercises 42-45 (and thus very little additional work should be required). These properties show that the $k$-metric really is a metric.

50. Show that $d_k(X, Y) \ge 0$ for any $X, Y$.

51. Show that $d_k(X, Y) = 0$ if and only if $\mathbb{P}(Y = X) = 1$ (so that $X$ and $Y$ are equivalent).

52. Show that $d_k(X, Y) \le d_k(X, Z) + d_k(Z, Y)$ for any $X, Y, Z$ (this is known as the **triangle inequality**).

Thus, the standard deviation is simply the 2-distance from $X$ to its mean:

$$\text{sd}(X) = d_2(X, \mathbb{E}(X)) = \sqrt{\mathbb{E}\left((X - \mathbb{E}(X))^2\right)}$$

and the variance is the square of this. More generally, the $k^{\text{th}}$ moment of $X$ about $a$ is simply the $k^{\text{th}}$ power of the $k$-distance from $X$ to $a$. The 2-distance is especially important for reasons that will become clear below and in the next section. This distance is also called the **root mean square distance**.

### Center and Spread Revisited

Measures of center and measures of spread are best thought of together, in the context of a **measure of distance**. For a random variable $X$, we first try to find the constants $t$ that are closest to $X$, as measured by the given distance; any such $t$ is a **measure of center** relative to the distance. The minimum distance itself is the corresponding **measure of spread**.

Let us apply this procedure to the 2-distance. Thus, we define the **root mean square** error function by

$$d_2(X, t) = \|X - t\|_2 = \sqrt{\mathbb{E}\left((X - t)^2\right)}, \quad t \in \mathbb{R}$$

53. Show that $d_2(X, t)$ is minimized when $t = \mathbb{E}(X)$ and that the minimum value is $\text{sd}(X)$.

a. Note that the minimum value of the function occurs at the same points as the minimum value of $\mathbb{E}\left((X - t)^2\right)$ (this is the **mean square** error function).
b. Expand $(X - t)^2$ and take expected values term by term. The resulting expression is a quadratic function of $t$.
c. Now use standard calculus.

The physical interpretation of this result is that the moment of inertia of the mass distribution of $X$ about $t$ is minimized when $t = \mu$, the center of mass.

54. In the histogram applet, construct a discrete distribution each of the types indicated below. Note the position and size of the mean ± standard deviation bar and the shape of the mean square error graph.

a. A uniform distribution.
b. A symmetric, unimodal distribution.
c. A unimodal distribution that is skewed right.
d. A unimodal distribution that is skewed left.
e. A symmetric bimodal distribution.
f. A U-shaped distribution.

Next, let us apply our procedure to the 1-distance. Thus, we define the **mean absolute error** function by

$$d_1(X, t) = \|X - t\|_1 = \mathbb{E}(|X - t|), \quad t \in \mathbb{R}$$

We will show that $d_1(X, t)$ is minimized when $t$ is any median of $X$. We start with a discrete case, because it's easier and has special interest.

55. Suppose that $X$ has a discrete distribution with values in a finite set $S$.

a. Show that $\mathbb{E}(|X - t|) = \mathbb{E}(t - X, X \le t) + \mathbb{E}(X - t, X > t)$
b. Show that $\mathbb{E}(|X - t|) = a_t\, t + b_t$, where $a_t = 2\,\mathbb{P}(X \le t) - 1$ and where $b_t = \mathbb{E}(X) - 2\,\mathbb{E}(X, X \le t)$
c. Argue that $\mathbb{E}(|X - t|)$ is a continuous, piecewise linear function of $t$, with corners at the values in $S$. That is, the function is a **linear spline**.
d. Let $m$ be the smallest median of $X$. Argue that if $t < m$ and $t \notin S$, then the slope of the linear piece at $t$ is negative.
e. Let $M$ be the largest median of $X$. Argue that if $t > M$ and $t \notin S$, then the slope of the linear piece at $t$ is positive.
f. Argue that if $t \in (m, M)$ then the slope of the linear piece at $t$ is 0.
g. Conclude that $\mathbb{E}(|X - t|)$ is minimized for every $t$ in the median interval $[m, M]$.

The last exercise shows that mean absolute error has a couple of basic deficiencies as a measure of error:

- The function may not be smooth (differentiable).
- The function may not have a unique minimizing value of $t$.

Indeed, when $X$ does not have a unique median, there is no compelling reason to choose one value in the median interval, as the measure of center, over any other value in the interval.

56. In the histogram applet, construct a distribution of each of the types indicated below. In each case, note the position and size of the boxplot and the shape of the mean absolute error graph.

a. A uniform distribution.
b. A symmetric, unimodal distribution.
c. A unimodal distribution that is skewed right.
d. A unimodal distribution that is skewed left.
e. A symmetric bimodal distribution
f. A U-shaped distribution.

57. Let $X$ be an indicator random variable with $\mathbb{P}(X = 1) = p$. Graph $d_1(X, t) = \mathbb{E}(|X - t|)$ as a function of $t \in \mathbb{R}$ in each of the cases below. In each case, find the minimum value of the function and the values of $t$ where the minimum occurs.

a. $p < \frac{1}{2}$
b. $p = \frac{1}{2}$
c. $p > \frac{1}{2}$

58. Suppose now that $X$ has a general distribution on $\mathbb{R}$. Show that $d_1(X, t)$ is minimized when $t$ is any median of $X$.

a. Suppose that $s < t$. Compute the expected value over the events $X \le s$, $s < X \le t$, and $X \ge t$, and then add and subtract appropriate stuff to show that
$$\mathbb{E}(|X - t|) = \mathbb{E}(|X - s|) + (t - s)(2\,\mathbb{P}(X \le s) - 1) + 2\,\mathbb{E}(t - X,\ s < X \le t)$$
b. Suppose that $t < s$. Use methods similar to those in (a) to show that
$$\mathbb{E}(|X - t|) = \mathbb{E}(|X - s|) + (t - s)(2\,\mathbb{P}(X < s) - 1) + 2\,\mathbb{E}(X - t,\ t \le X < s)$$
c. Note that the last terms on the right in the equations in (a) and (b) are nonnegative. If we take $s$ to be a median of $X$, then the middle terms on the right in the equations in (a) and (b) are also nonnegative.
d. Conclude that if $s$ is a median of $X$ and $t$ is any other number then $\mathbb{E}(|X - t|) \ge \mathbb{E}(|X - s|)$.

## Convergence

Whenever we have a measure of distance, we automatically have a criterion for convergence. Let $X_n$, $n \in \{1, 2, ...\}$ and $X$ be real-valued random variables defined on the same sample space (that is, defined for the same random experiment). We say that $X_n \to X$ as $n \to \infty$ in **$k^{\text{th}}$ mean** if

$$d_k(X_n, X) \to 0 \text{ as } n \to \infty$$

or equivalently

$$\mathbb{E}\left(|X_n - X|^k\right) \to 0 \text{ as } n \to \infty$$

When $k = 1$, we simply say that $X_n \to X$ as $n \to \infty$ **in mean**; when $k = 2$, we say that $X_n \to X$ as $n \to \infty$ **in mean square**. These are the most important special cases.

59. Use Lyapunov's inequality to show that if $j < k$, then $X_n \to X$ as $n \to \infty$ in $k^{\text{th}}$ mean implies $X_n \to X$ as $n \to \infty$ in $j^{\text{th}}$ mean.

Our next sequence of exercises shows that convergence in mean is stronger than convergence in probability.

60. Use Markov's inequality to show that if $X_n \to X$ as $n \to \infty$ in mean, then $X_n \to X$ as $n \to \infty$ in probability.

The converse is not true. Moreover, convergence with probability 1 does not imply convergence in $k^{\text{th}}$ mean and convergence in $k^{\text{th}}$ mean does not imply convergence with probability 1. The next two exercises give some counterexamples.

61. Suppose that $(X_1, X_2, ...)$ is a sequence of independent random variables with

$$\mathbb{P}\left(X_n = n^3\right) = \frac{1}{n^2}, \; \mathbb{P}(X_n = 0) = 1 - \frac{1}{n^2}, \quad n \in \mathbb{N}_+$$

a. Use the first Borel-Cantelli lemma to show that $X_n \to 0$ as $n \to \infty$ with probability 1.
b. Show that $X_n \to 0$ as $n \to \infty$ in probability.
c. Show that $\mathbb{E}(X_n) \to \infty$ as $n \to \infty$.

62. Suppose that $(X_1, X_2, ...)$ is a sequence of independent indicator random variables with

$$\mathbb{P}(X_n = 1) = \frac{1}{n}, \; \mathbb{P}(X_n = 0) = 1 - \frac{1}{n}, \quad n \in \mathbb{N}_+$$

a. Use the second Borel-Cantelli lemma to show that $\mathbb{P}(X_n = 0 \text{ for infinitely many } n) = 1$.
b. Use the second Borel-Cantelli lemma to show that $\mathbb{P}(X_n = 1 \text{ for infinitely many } n) = 1$.
c. Show that $\mathbb{P}(X_n \text{ does not converge as } n \to \infty) = 1$
d. Show that $X_n \to 0$ as $n \to \infty$ in $k^{\text{th}}$ mean for every $k \ge 1$.

The implications in the various modes of convergence are shown below; no other implications hold in general.

- Convergence with probability 1 implies convergence in probability.
- Convergence in $k^{\text{th}}$ mean implies convergence in $j^{\text{th}}$ mean if $j \le k$.

- Convergence in $k^{\text{th}}$ mean implies convergence in probability.
- Convergence in probability implies convergence in distribution.

## Related Topics

For a related statistical topic, see the section on the Sample Variance in the chapter on Random Samples. The variance of a sum of random variables is best understood in terms of a related concept known as covariance, that will be studied in detail in the next section.

---

Virtual Laboratories > 4. Expected Value > 1 2 3 4 5 6
Contents | Applets | Data Sets | Biographies | External Resources | Keywords | Feedback | ©