

3. The Sample Variance

Descriptive Statistics

Once again, our first discussion is from a purely descriptive point of view, and thus without reference to any underlying probability distributions. First recall the basic model of descriptive statistics: we have a population of objects of interest, and we have various measurements (variables) that we make on these objects. We select objects from the sample and record the variables for these objects; these become our data.

Variance and Standard Deviation

Suppose that $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is a sample of size n from a real-valued variable x . Recall that the [sample mean](#) is

$$m = \frac{1}{n} \sum_{i=1}^n x_i$$

and is the most important measure of the center of the data set. The **sample variance** is defined to be

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2$$

If we need to indicate the dependence on the data set \mathbf{x} , we write $s^2(\mathbf{x})$. The difference $x_i - m$ is the **deviation** of the i^{th} data value x_i from the mean m of the data set. Thus, the variance is the **mean square deviation** and is a measure of the spread of the data set with respect to the mean. The reason for dividing by $n - 1$ rather than n is best understood in terms of the [inferential point of view](#) that we discuss below; this definition makes the sample variance an unbiased estimator of the distribution variance. However, the reason for the averaging can also be understood in terms of a related concept.

1. Show that $\sum_{i=1}^n (x_i - m) = 0$.

Thus, if we know $n - 1$ of the deviations, we can compute the last one. This means that there are only $n - 1$ freely varying deviations, that is to say, $n - 1$ **degrees of freedom** in the set of deviations. In the definition of sample variance, we average the squared deviations, not by dividing by the number of terms, but rather by dividing by the number of degrees of freedom in those terms. However, this argument notwithstanding, it would be reasonable, *from a purely descriptive point of view*, to divide by n in the definition of the sample variance. Moreover, when n is sufficiently large, it does not make much difference whether we divide by n or by $n - 1$.

In any event, the square root s of the sample variance s^2 is the sample standard deviation. It is the **root mean square deviation** and is also a measure of the spread of the data with respect to the mean. Both measures of spread are important. Variance has nicer mathematical properties, but its physical unit is the square of the unit of x . For example, if the underlying variable x is the height of a person in inches, the variance is in square inches. On the other hand, the standard

deviation has the same physical unit as the original variable, but its mathematical properties are not as nice.

Measures of Center and Spread

Measures of center and measures of spread are best thought of together, in the context of an **error function**. The error function measures how well a single number a represents the entire data set x . The values of a (if they exist) that minimize the error functions are our measures of center; the minimum value of the error function is the corresponding measure of spread. Of course, we hope for a *single* value of a that minimizes the error function, so that we have a unique measure of center.

Let's apply this procedure to the **mean square deviation** function defined by

$$\text{msd}(a) = \frac{1}{n-1} \sum_{i=1}^n (x_i - a)^2$$

Minimizing msd is a standard problem in calculus.

2. Note that the graph of msd is a parabola opening upward. Show that

- msd is minimized when $a = m$, the sample mean.
- The minimum value of msd is s^2 , the sample variance.

Trivially, if we defined the mean square error function by dividing by n rather than $n - 1$, then the minimum value would still occur at m , the sample mean, but the minimum value would be the alternate version of the sample variance in which we divide by n . On the other hand, if we were to use the **root mean square deviation** function $\text{rmsd}(a) = \sqrt{\text{msd}(a)}$, then because the square root function is strictly increasing on $[0, \infty)$, the minimum value would again occur at m , the sample mean, but the minimum value would be s , the sample standard deviation. The important point is that with all of these error functions, the unique measure of center is the sample mean, and the corresponding measures of spread are the various ones that we are studying.

Next, let's apply our procedure to the **mean absolute deviation** function defined by

$$\text{mad}(a) = \frac{1}{n-1} \sum_{i=1}^n |x_i - a|$$

3. Note that

- mad is a continuous function.
- The graph of mad consists of lines.
- The slope of the line at a depends on where a is in the data set x

Mathematically, mad has some problems as an error function. First, the function will not be smooth (differentiable) at points where two lines of different slopes meet. More importantly, the values that minimize mad may occupy an entire interval, thus leaving us without a unique measure of center. The [histogram exercises](#) below will show you that these

pathologies can really happen. It turns out that mad is minimized at any point in the **median interval** of the data set \mathbf{x} . Thus, the medians are the natural measures of center associated with mad as a measure of error, in the same way that the sample mean is the measure of center associated with the msd as a measure of error.

Properties

In this section, we establish some essential properties of the sample variance and standard deviation. First, the following alternate formula for the sample variance is better for computational purposes, and for certain theoretical purposes as well.

4. Show that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} m^2$$

If we let $\mathbf{x}^2 = (x_1^2, x_2^2, \dots, x_n^2)$ denote the sample from the variable x^2 , then the computational formula in the last exercise can be written succinctly as

$$s^2(\mathbf{x}) = \frac{n}{n-1} (m(\mathbf{x}^2) - m^2(\mathbf{x}))$$

5. Show that

- $s^2 \geq 0$
- $s^2 = 0$ is and only if $x_i = m$ for each $i \in \{1, 2, \dots, n\}$, so that the data set is constant.

6. Suppose that c is a constant. Show that

- $s^2(c \mathbf{x}) = c^2 s^2(\mathbf{x})$
- $s(c \mathbf{x}) = |c| s(\mathbf{x})$

7. Suppose that \mathbf{c} is a constant vector of size n . Show that $s^2(\mathbf{x} + \mathbf{c}) = s^2(\mathbf{x})$

Thus, multiplying the data by a positive constant c changes the standard deviation by a factor of c also. This operation can always be thought of a **change of scale** in the underlying variable x . On the other hand, adding a constant to the data does not change the standard deviation. This operation can be thought of as a **change of location**. Now, for $i \in \{1, 2, \dots, n\}$, let

$$z_i = \frac{x_i - m}{s}$$

for $i \in \{1, 2, \dots, n\}$. The number z_i is the **standard score** associated with x_i . Note that since x_i , m , and s have the same physical units, the standard score z_i is **dimensionless** (that is, has no physical units); it measures the directed distance from the mean m to the data value x_i in standard deviations.

8. Show that the sample of standard scores $\mathbf{z} = (z_1, z_2, \dots, z_n)$ has mean 0 and variance 1. That is,

- a. $m(\mathbf{z}) = 0$
- b. $s^2(\mathbf{z}) = 1$

Approximating the Variance

Suppose that instead of the actual data \mathbf{x} , we have a **frequency distribution** with classes (intervals) $\{A_j : j \in J\}$, class marks (midpoints of the intervals) $\{t_j : j \in J\}$, and frequencies $\{n_j : j \in J\}$. Recall that the relative frequency of class A_j is $p_j = \frac{n_j}{n}$. In this case, approximate values of the sample mean and variance are, respectively,

$$m = \frac{1}{n} \sum_{j \in J} n_j t_j = \sum_{j \in J} p_j t_j$$
$$s^2 = \frac{1}{n-1} \sum_{j \in J} n_j (t_j - m)^2 = \frac{n}{n-1} \sum_{j \in J} p_j (t_j - m)^2$$

These approximations are based on the hope that the data values in each class are well represented by the class mark.

Random Samples

We now turn our attention to the more interesting inferential setting where the data are random, generated by an underlying probability distribution. Thus, suppose that we have a basic **random experiment**, and that X is a real-valued **random variable** for the experiment with **mean** μ and **standard deviation** σ . We will need some other distribution moments as well. For $k \in \mathbb{N}_+$, let $d_k = \mathbb{E}((X - \mu)^k)$ denote the k^{th} moment about the mean. In particular, note that $d_0 = 1$, $d_1 = 0$, and $d_2 = \sigma^2$. We assume that $d_4 < \infty$.

We repeat the basic experiment n times to form a new, compound experiment, with a sequence of independent random variables $\mathbf{X} = (X_1, X_2, \dots, X_n)$, each with the same distribution as X . In statistical terms, \mathbf{X} is a **random sample** of size n from the distribution of X . All of the statistics defined above in the section on **descriptive statistics** make sense for \mathbf{X} , of course, but now these statistics are random variables. We will use the notation established above, except for the usual convention of denoting random variables by capital letters. Finally, note that the deterministic properties and relations established in the section on descriptive statistics still hold.

In addition to being a measure of the center of the data \mathbf{X} , the **sample mean**

$$M = \frac{1}{n} \sum_{i=1}^n X_i$$

is a natural estimator of the distribution mean μ . In this section, we will derive statistics that are natural estimators of the distribution variance σ^2 . The statistics that we will derive are different, depending on whether μ is known or unknown; for this reason, μ is referred to as a **nuisance parameter** for the problem of estimating σ^2 .

A Special Sample Variance

First we will assume that μ is known. Although this is almost always an artificial assumption, it is a nice place to start because the analysis is relatively easy. Let

$$W^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

9. Note that W^2 is the sample mean for a random sample of size n from the distribution of $(X - \mu)^2$. Thus, conclude that

- $\mathbb{E}(W^2) = \sigma^2$.
- $\text{var}(W^2) = \frac{1}{n} (d_4 - \sigma^4)$.
- $W^2 \rightarrow \sigma^2$ as $n \rightarrow \infty$ with probability 1.

In particular part (a) means that W^2 is an **unbiased** estimator of σ^2 . The square root of the special sample variance is a special version of the **sample standard deviation**, denoted W .

10. Use **Jensen's inequality** to show that $\mathbb{E}(W) \leq \sigma$. Thus, W is a **biased** estimator that tends to underestimate σ .

11. Use **basic properties of covariance** to show that $\text{cov}(M, W^2) = \frac{d_3}{n}$.

It follows that the sample mean and the special sample variance are uncorrelated if $d_3 = 0$ and are **asymptotically uncorrelated** in any case.

The Standard Sample Variance

Consider now the more realistic case in which μ is unknown. In this case, a natural approach is to average, in some sense, the squared deviations $(X_i - M)^2$ over $i \in \{1, 2, \dots, n\}$. It might seem that we should average by dividing by n .

However, another approach is to divide by whatever constant would give us an unbiased estimator of σ^2 . This constant turns out to be $n - 1$, leading to the **standard sample variance**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2$$

12. Show that $\mathbb{E}(S^2) = \sigma^2$

- Recall that $\mathbb{E}(M) = \mu$
- Recall that $\text{var}(M) = \frac{\sigma^2}{n}$.
- Now use **Exercise 4**.

Of course, the square root of the sample variance is the **sample standard deviation**, denoted S .

13. Use Jensen's inequality to show that $\mathbb{E}(S) \leq \sigma$. Thus, S is a biased estimator than tends to underestimate σ .

14. Use Exercise 4 and the (strong) law of large numbers to show that $S^2 \rightarrow \sigma^2$ as $n \rightarrow \infty$ with probability 1.

Higher Moments of the Sample Variance

In this subsection, we will derive formulas for the covariance between the sample mean and the sample variance and for the variance of the sample variance. The first tool that we need is a relationship between the standard sample variance, the special sample variance, and the sample mean.

15. Show that $S^2 = \frac{n}{n-1} (W^2 - (M - \mu)^2)$

Next we need an expansion of the second term in the last exercise.

16. Show that

$$(M - \mu)^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (X_i - \mu)(X_j - \mu)$$

Our first main result is the covariance between the sample mean and the sample variance.

17. Show that $\text{cov}(M, S^2) = \frac{d_3}{n}$.

a. Use Exercise 15 to show that $\text{cov}(S^2, M) = \frac{n}{n-1} (\text{cov}(W^2, M) - \text{cov}((M - \mu)^2, M))$

b. We know the first covariance on the right from Exercise 11.

c. Note that $\text{cov}((M - \mu)^2, M) = \text{cov}((M - \mu)^2, M - \mu)$

d. Using Exercise 16 and properties of covariance, show that

$$\text{cov}((M - \mu)^2, M - \mu) = \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \text{cov}((X_i - \mu)(X_j - \mu), X_k - \mu)$$

e. In part (d), the covariances are 0 except when $i = j = k$, and in this case the covariance is d_3 . There are n such terms.

f. Now put the pieces together to derive the result.

In particular, note that $\text{cov}(M, S^2) = \text{cov}(M, W^2)$. Again, the sample mean and variance are uncorrelated if $d_3 = 0$, and asymptotically uncorrelated otherwise. Our last major result is the variance of the sample variance.

18. Show that $\text{var}(S^2) = \frac{1}{n} \left(d_4 - \frac{n-3}{n-1} \sigma^4 \right)$.

a. Use Exercise 15 to show that $\text{var}(S^2) = \left(\frac{n}{n-1} \right)^2 \left(\text{var}(W^2) - 2 \text{cov}(W^2, (M - \mu)^2) + \text{var}((M - \mu)^2) \right)$

b. We know $\text{var}(W^2)$ from Exercise 9.

c. Use Exercise 16 to show that

$$\text{cov}(W^2, (M - \mu)^2) = \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \text{cov}\left((X_i - \mu)^2, (X_j - \mu)(X_k - \mu)\right).$$

d. In part (c), show that the covariances are all 0 except when $i = j = k$, and in this case the covariance is $d_4 - \sigma^4$.

There are n such terms.

e. Use Exercise 16 to show that

$$\text{var}((M - \mu)^2) = \frac{1}{n^4} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n \text{cov}\left((X_i - \mu)(X_j - \mu), (X_k - \mu)(X_l - \mu)\right).$$

f. In part (e), show that the covariance is $d_4 - \sigma^4$ if $i = j = k = l$. There are n such terms.

g. In part (e), show that the covariance is σ^4 if $i \neq j$ and $\{i, j\} = \{k, l\}$. There are $2n(n-1)$ such terms.

h. In part (e), show that the covariance is 0 in all other cases.

i. Now put the pieces together to derive the result for the variance.

❖ 19. Show that $\text{var}(S^2) > \text{var}(W^2)$. Does this seem reasonable?

❖ 20. Show that $\text{var}(S^2) \rightarrow 0$ as $n \rightarrow \infty$. This means that S^2 is a **consistent** estimator of σ^2 .

As a corollary to the other results in this subsection, we can compute the covariance between the special sample variance and the standard one. Curiously, it's the same as the variance of the special sample variance.

❖ 21. Use Exercise 9(b), Exercise 15, and Exercise 18(c) to show that $\text{cov}(S^2, W^2) = \frac{1}{n}(d_4 - \sigma^4)$.

Exercises

Simulation Exercises

Many of the applets in this project are simulations of experiments with a basic random variable of interest. When you run the simulation, you are performing independent replications of the experiment. In most cases, the applet displays the standard deviation of the distribution, both numerically in a table and graphically as the radius of the blue, horizontal bar in the graph box. When you run the simulation, the sample standard deviation is also displayed numerically in the table and graphically as the radius of the red horizontal bar in the graph box.

❖ 22. In the **binomial coin experiment**, the random variable is the number of heads. For various values of the parameters n (the number of coins) and p (the probability of heads), run the simulation 1000 times and note the apparent convergence of the sample standard deviation to the distribution standard deviation.

❖ 23. In the simulation of the **matching experiment**, the random variable is the number of matches. For selected values of n (the number of balls), run the simulation 1000 times and note the apparent convergence of the sample standard deviation to the distribution standard deviation.

❖ 24. Run the simulation of the **gamma experiment** 1000 times for various values of the rate parameter r and the shape parameter k . Note the apparent convergence of the sample standard deviation to the distribution standard deviation.

Histogram and MAD Exercises

25. In the [histogram applet](#), select mean and standard deviation. Set the class width to 0.1 and construct a distribution with at least 30 values of each of the types indicated below. Then increase the class width to each of the other four values. As you perform these operations, note the position and size of the mean-standard deviation bar.

- A uniform distribution.
- A symmetric, unimodal distribution.
- A unimodal distribution that is skewed right.
- A unimodal distribution that is skewed left.
- A symmetric bimodal distribution.
- A u -shaped distribution.

26. In the [histogram applet](#), construct a distribution with 20 points that has the largest possible standard deviation.

27. In the [histogram error function applet](#), select mean square error. As you add points, note the shape of the graph of the error function, the value that minimizes the function, and the minimum value of the function.

28. In the [histogram error function applet](#), select mean absolute error. As you add points, note the shape of the graph of the error function, the values that minimizes the function, and the minimum value of the function.

29. Suppose that our data set is $(2, 5, 2)$. Explicitly give mad as a piecewise function and sketch its graph. Note that

- All values of $a \in [2, 5]$ minimize mad.
- mad is not differentiable at $a \in \{2, 5\}$.

30. Suppose that our data set is $(3, 5, 1)$. Explicitly give mad as a piecewise function and sketch its graph. Note that

- mad is minimized at $a = 3$
- mad is not differentiable at $a \in \{1, 3, 5\}$.

Data Analysis Exercises

31. Compute the sample mean and standard deviation and plot a density histogram for [Michelson's velocity of light data](#).



32. Compute the sample mean and standard deviation and plot a density histogram for [Cavendish's density of the earth data](#).



33. Consider the [M&M data](#).

- Compute the sample mean and standard deviation and plot a frequency histogram for the total number of candies.

b. Compute the sample mean and standard deviation and plot a density histogram for the net weight.



34. Consider **Fisher's iris data**. Compute the sample mean and standard deviation and plot a density histogram of petal length, with the restrictions given below. Compare the results.

- a. All cases
- b. Each species individually



35. Consider the **Cicada data**. Plot a density histogram and compute the sample mean and sample standard deviation for body weight, with the restrictions given below. Compare the results.

- a. All cases
- b. Each species individually
- c. Male and female individually



36. Consider the **Pearson's height data**. Plot a density histogram and compute the sample mean and sample standard deviation for

- a. the height of the father
- b. the height of the son



37. Consider the **first Challenger data set**. Plot a density histogram and compute the sample mean and sample standard deviation for the erosion variable. For the histogram, use the classes $[0, 5)$, $[5, 40)$, $[40, 50)$, $[50, 60)$.

