Virtual Laboratories > 6. Random Samples > 1 2 3 4 5 6 7

3. The Sample Variance

Preliminaries

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. Additionally, let

$$d_k = \mathbb{E}\left((X - \mu)^k\right), \quad k \in \mathbb{N}$$

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 $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, each with the same distribution as X. In statistical terms, $\boldsymbol{X}$ is a random sample of size n from the distribution of X. Recall that the sample mean

$$M(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a natural measure of the center of the data and a natural estimator of the distribution mean μ. In this section, we will derive statistics that are natural measures of the dispersion of the data and 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

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

Properties

1. Show that W^2 is the sample mean for a random sample of size n from the distribution of $(X - \mu)^2$.

2. Use the result of Exercise 1 to show that

   a. $\mathbb{E}(W^2) = \sigma^2$.

b. $\text{var}\left(W^2\right) = \frac{1}{n}\left(d_4 - \sigma^4\right)$.

c. $W^2 \to \sigma^2$ as $n \to \infty$ with probability 1.

In particular 2 (a) means that W^2 is an **unbiased** estimator of σ^2.

3. Use basic properties of covariance to show that $\text{cov}\left(M,\ W^2\right) = \frac{d_3}{n}$. 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 square root of the special sample variance is a special version of the **sample standard deviation**, denoted $W(\boldsymbol{X})$.

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

5. Show that if c is a constant then $W^2\left(c\,\boldsymbol{X}\right) = c^2\,W^2\left(\boldsymbol{X}\right)$

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, $(X_i - M)^2$ over $i \in \{1, 2, ..., n\}$. 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.

6. Use basic algebra to show that

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

7. Use the result in Exercise 6 and basic properties of expected value to show that

$$\mathbb{E}\left(\sum_{i=1}^{n} (X_i - M)^2\right) = (n-1)\,\sigma^2$$

From Exercise 7, the random variable

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

is an unbiased estimator of σ^2; it is called the **sample variance**. As a practical matter, when n is large, it makes little difference whether we divide by n or $n - 1$.

Basic Properties

The following alternate formula follows immediately from Exercise 6, and it is better for some purposes.

8. Show that

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

9. Use the formula in the previous exercise and the (strong) law of large numbers to show that $S^2 \to \sigma^2$ as $n \to \infty$ with probability 1.

10. Show that if c is a constant then $S^2(c\,\boldsymbol{X}) = c^2\, S^2(\boldsymbol{X})$

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

The square root of the sample variance is the **sample standard deviation**, denoted $S(\boldsymbol{X})$.

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

Moments

In this section we will derive formulas for the variance of the sample variance and the covariance between the sample mean and the sample variance. Our first series of exercises will show that

$$\operatorname{var}\left(S^2\right) = \frac{1}{n}\left(d_4 - \frac{n-3}{n-1}\sigma^4\right)$$

13. Verify the following result. *Hint*: Start with the expression on the right. Expand the term $\left(X_i - X_j\right)^2$, and take the sums term by term.

$$S^2 = \frac{1}{2\,n\,(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(X_i - X_j\right)^2$$

It follows that $\operatorname{var}\left(S^2\right)$ is the sum of all of the pairwise covariances of the terms in the expansion of Exercise 13.

14. Suppose that $i \ne j$, Verify the following results. (*Hint*: In $\mathbb{E}\left(\left(X_i - X_j\right)^m\right)$, add and subtract μ, and then expand and use independence.)

a. $\mathbb{E}\left(\left(X_i - X_j\right)^2\right) = 2\,\sigma^2$
b. $\mathbb{E}\left(\left(X_i - X_j\right)^4\right) = 2\,d^4 + 6\,\sigma^4$

15. Finally, derive the formula for $\operatorname{var}\left(S^2\right)$ by showing that

a. $\operatorname{cov}\left(\left(X_i - X_j\right)^2, (X_k - X_l)^2\right) = 0$ if $i = j$ or $k = l$ or i, j, k, l are distinct.
b. $\operatorname{cov}\left(\left(X_i - X_j\right)^2, \left(X_i - X_j\right)^2\right) = 2\,d_4 + 2\,\sigma^4$ if $i \ne j$, and there are $2\,n\,(n-1)$ such terms in the sum of

covariances.

c. $\operatorname{cov}\left(\left(X_i - X_j\right)^2, \left(X_k - X_j\right)^2\right) = d_4 - \sigma^4$ if i, j, k are distinct, and there are $4\,n\,(n-1)\,(n-2)$ such terms in the sum of covariances.

16. Show that $\operatorname{var}\left(S^2\right) > \operatorname{var}\left(W^2\right)$. Does this seem reasonable?

17. Show that $\operatorname{var}\left(S^2\right) \to 0$ as $n \to \infty$.

18. Use similar techniques to show that $\operatorname{cov}\left(M, S^2\right) = \frac{d_3}{n}$. In particular, note that $\operatorname{cov}\left(M, S^2\right) = \operatorname{cov}\left(M, W^2\right)$. Again, the sample mean and variance are uncorrelated if $d_3 = 0$, and asymptotically uncorrelated otherwise.

Examples and Special Cases

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, sample standard deviation is also displayed numerically in the table and graphically as the radius of the red horizontal bar in the graph box.

19. In the binomial coin experiment, the random variable is the number of heads. Run the simulation 1000 times updating every 10 runs and note the apparent convergence of the sample standard deviation to the distribution standard deviation.

20. In the simulation of the matching experiment, the random variable is the number of matches. Run the simulation 1000 times updating every 10 runs and note the apparent convergence of the sample standard deviation to the distribution standard deviation.

21. Run the simulation of the exponential experiment 1000 times with an update frequency of 10. Note the apparent convergence of the sample standard deviation to the distribution standard deviation.

Data Analysis Exercises

The sample mean and standard deviation are often computed in exploratory data analysis, as measures of the center and spread of the data, respectively.

22. Compute the sample mean and standard deviation for Michelson's velocity of light data.

?

23. Compute the sample mean and standard deviation for Cavendish's density of the earth data.

?

24. Compute the sample mean and standard deviation of the net weight in the M&M data.

25. Compute the sample mean and standard deviation of the petal length variable for the following cases in Fisher's iris data. Compare the results.

a. All cases
b. Setosa only
c. Versicolor only
d. Verginica only

Interval Data

Suppose that instead of the actual data, we have a frequency distribution with classes $(A_1, A_2, ..., A_k)$, class marks $(x_1, x_2, ..., x_k)$, and frequencies $(n_1, n_2, ..., n_k)$. Thus,

$$n_j = \#\left(\left\{i \in \{1, 2, ..., n\} : X_i \in A_j\right\}\right), \quad j \in \{1, 2, ..., k\}$$

In this case, approximate values of the sample mean and variance are, respectively,

$$m = \frac{1}{n} \sum_{j=1}^{k} n_j\, x_j, \quad s^2 = \frac{1}{n-1} \sum_{j=1}^{k} n_j \left(x_j - m\right)^2$$

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

26. In the interactive histogram, select mean and standard deviation. Set the class width to 0.1 and construct a frequency distribution with at least 6 nonempty classes and at least 10 values. Compute the mean, variance, and standard deviation by hand, and verify that you get the same results as the applet.

27. In the interactive histogram, 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. 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.

28. In the interactive histogram, construct a distribution that has the largest possible standard deviation.

⚄ 29. Based on your answer to Exercise 28, characterize the distributions (on a fixed interval [*a*, *b*]) that have the largest possible standard deviation.

Virtual Laboratories > 6. Random Samples > 1 2 3 4 5 6 7
Contents | Applets | Data Sets | Biographies | External Resources | Keywords | Feedback | ©