

2. The Sample Mean and the Law of Large Numbers

The Sample Mean

Suppose that we have a basic [random experiment](#) and that X is a real-valued [random variable](#) for the basic experiment. Now suppose we perform n independent replications of the basic experiment. This defines 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 . Recall that in statistical terms, \mathbf{X} is a [random sample](#) of size n from the distribution of X . The [sample mean](#) is simply the average of the variables in the sample:

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

The sample mean is a function of the random sample and thus is a *statistic*. Like any statistic, the sample mean is itself a random variable with a distribution, mean, and variance of its own. Many times, the distribution mean is unknown and the sample mean is used as an [estimator](#) of this unknown parameter. When the underlying sample is understood, we will usually suppress it from the notation.

1. In the [dice experiment](#), the dice scores form a random sample from the specified die distribution. Select the average random variable; this variable is the sample mean for the sample of dice scores. For each die distribution, start with 1 die and increase the sample size n . Note the shape and location of the probability density function of the sample mean at each stage. For selected values of n , run the simulation 1000 times with an update frequency of 10. Note the apparent convergence of the empirical density function to the true density function.

Moments

Suppose that X has [mean](#) μ and [standard deviation](#) σ (assumed finite).

2. Use basic properties of expected value to show that $\mathbb{E}(M) = \mu$.

Exercise 2 shows that the sample mean M is an [unbiased](#) estimator of the distribution mean μ . Therefore, the variance of M is the [mean square error](#), when M is used as an estimator of μ .

3. Use basic properties of variance to show that $\text{var}(M) = \frac{\sigma^2}{n}$.

From Exercise 3, the variance of the sample mean is an increasing function of the distribution variance and a decreasing function of the sample size. Both of these make intuitive sense if we think of the sample mean as an estimator of the distribution mean.

4. In the **dice experiment**, select the average random variable, which as noted before, is the sample mean for the sample of dice scores. For each die distribution, start with 1 die and increase the sample size n . Note that the mean of the sample mean stays the same, but the standard deviation of the sample mean decreases (as we now know, in inverse proportion to the square root of the sample size). For selected values of n , run the simulation 1000 times, updating every 10 runs. Note the apparent convergence of the empirical moments of the sample mean to the true moments.

5. Compute the sample mean of the petal width variable for the following cases in **Fisher's iris data**. Compare the results.

- All cases
- Setosa only
- Versicolor only
- Verginica only

Linearity

Computing the sample mean is a linear operation.

6. Suppose that X and Y are random samples of size n (defined on the same probability space), and that c is a constant. Show that

- $M(X + Y) = M(X) + M(Y)$
- $M(cX) = cM(X)$

The Law of Large Numbers

The **law of large numbers** states that the sample mean converges to the distribution mean as the sample size increases, and is one of the fundamental theorems of probability. There are different versions of the law, depending on the *mode* of convergence.

Suppose again that X is a real-valued random variable for our basic experiment, with mean μ and standard deviation σ (assumed finite). We repeat the basic experiment indefinitely to create a new, compound experiment with an infinite sequence of independent random variables (X_1, X_2, \dots) , each with the same distribution as X . In statistical terms, we are **sampling** from the distribution of X . For each n , let M_n denote the sample mean of the first n sample variables:

$$M_n = M(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

Weak Laws

From [Exercise 3](#), note that $\text{var}(M_n) = \mathbb{E}((M_n - \mu)^2) \rightarrow 0$ as $n \rightarrow \infty$. This means that $M_n \rightarrow \mu$ as $n \rightarrow \infty$ in

mean square.

7. Use [Chebyshev's inequality](#) to show that for any $\varepsilon > 0$.

$$\mathbb{P}(|M_n - \mu| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

This means that $M_n \rightarrow \mu$ as $n \rightarrow \infty$ [in probability](#). Recall that in general, convergence in mean square implies convergence in probability. The convergence of the sample mean to the distribution mean in mean square and in probability are known as **weak laws of large numbers**.

The Strong Law

The **strong law of large numbers** states that the sample mean M_n converges to the distribution mean μ [with probability 1](#). As the name suggests, this is a much stronger result than the weak law. That is, the strong law of large numbers states that

$$\mathbb{P}(M_n \rightarrow \mu \text{ as } n \rightarrow \infty) = 1$$

The following exercises sketch the proof of the strong law of large numbers. First, let $Y_n = \sum_{i=1}^n X_i$ so that $M_n = \frac{Y_n}{n}$.

8. Use Chebyshev's inequality to show that for every $n \in \mathbb{N}_+$ and every $\varepsilon > 0$,

$$\mathbb{P}(|M_{n^2} - \mu| > \varepsilon) \leq \frac{\sigma^2}{n^2 \varepsilon^2}$$

9. Use the result of the previous exercise and the [first Borel-Cantelli lemma](#) to show that for every $\varepsilon > 0$,

$$\mathbb{P}(|M_{n^2} - \mu| > \varepsilon \text{ for infinitely many } n \in \mathbb{N}_+) = 0$$

10. Use the result of the previous exercise and [Boole's inequality](#) to show that

$$\mathbb{P}(\text{There exists a rational } \varepsilon > 0 \text{ such that } |M_{n^2} - \mu| > \varepsilon \text{ for infinitely many } n \in \mathbb{N}_+) = 0$$

11. Conclude from the previous exercise that $\mathbb{P}(M_{n^2} \rightarrow \mu \text{ as } n \rightarrow \infty) = 1$

Suppose temporarily that the underlying sampling variable is nonnegative: $\mathbb{P}(X \geq 0) = 1$

12. Show that with probability 1, Y_n is increasing in n .

For $n \in \mathbb{N}_+$, let k_n be the unique positive integer such that $k_n^2 \leq n < (k_n + 1)^2$.

13. Use the result of the previous exercise and algebra to show that, with probability 1,

$$\frac{Y_{k_n^2}}{(k_n + 1)^2} \leq \frac{Y_n}{n} \leq \frac{Y_{(k_n + 1)^2}}{k_n^2}$$

14. Use [Exercise 11](#) and [Exercise 13](#) and the “squeeze theorem” for limits to conclude the strong law of

large numbers for a nonnegative variable: $\mathbb{P}(M_n \rightarrow \mu \text{ as } n \rightarrow \infty) = 1$

We now relax the condition that the underlying sampling variable X be nonnegative, and we recall the definitions of the **positive and negative parts** a real number x :

$$x^+ = \max \{x, 0\}, \quad x^- = \max \{-x, 0\}$$

Recall also that $x^+ \geq 0$, $x^- \geq 0$, $x = x^+ - x^-$, and $|x| = x^+ + x^-$.

15. Use [Exercise 6](#) and [Exercise 14](#) to show that with probability 1,

- $\frac{1}{n} \sum_{i=1}^n X_i^+ \rightarrow \mathbb{E}(X^+)$ as $n \rightarrow \infty$
- $\frac{1}{n} \sum_{i=1}^n X_i^- \rightarrow \mathbb{E}(X^-)$ as $n \rightarrow \infty$

16. Finally conclude that $M_n \rightarrow \mu$ as $n \rightarrow \infty$ with probability 1.

The proof of the strong law of large numbers sketched above requires that the variance of the sampling distribution be finite (note that this is critical in [Exercise 8](#) and [Exercise 9](#)). However, there are better proofs that only require that $\mathbb{E}(|X|) < \infty$. See, for example, *Probability and Measure* by Patrick Billingsley

Simulation Exercises

17. In the [dice experiment](#), recall again that the dice scores form a random sample from the specified die distribution. Select the average random variable, which is the sample mean of the sample of dice scores. For each die distribution, start with 1 die and increase the sample size n . Note how the distribution of the sample mean begins to resemble a point mass distribution. Run the simulation 1000 times, updating every 10 runs. Note the apparent convergence of the empirical density of the sample mean to the true density.

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 mean of the distribution numerically in a table and graphically as the center of the blue horizontal bar in the graph box. When you run the simulation, sample mean is also displayed numerically in the table and graphically as the center of the red horizontal bar in the graph box.

18. In the simulation of the [binomial coin experiment](#), select the number of heads. For selected values of the parameters, run the simulation 1000 times updating every 10 runs and note the apparent convergence of the sample mean to the distribution mean.

19. In the simulation of the [matching experiment](#), the random variable is the number of matches. For selected values of the parameter, run the simulation 1000 times updating every 10 runs and note the apparent convergence of the sample mean to the distribution mean.

20. In the simulation of the [gamma experiment](#), the random variable represents a random arrival time. For selected values of the parameters, run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the sample mean to the distribution mean.

Extensions and Special Cases

Random samples and their sample means are ubiquitous in probability and statistics. In this subsection, we will see how sample means can be used to estimate probabilities, density functions, and distribution functions.

Relative Frequency

Suppose that X is a random variable for a basic experiment, taking values in a space S . Note that X might be the outcome variable for the entire experiment, in which case S would be the sample space. In any event, S is a general space so X may be vector-valued. Recall that the **distribution** of X is the probability measure on S given by

$$A \mapsto \mathbb{P}(X \in A)$$

Now suppose that we repeat the basic experiment n indefinitely to form a random sample (X_1, X_2, \dots, X_n) of size n from the distribution of X . The **empirical distribution** of X for this sample is defined by

$$P_n(A) = \frac{1}{n} \#\{i \in \{1, 2, \dots, n\} : X_i \in A\}, \quad A \subseteq S$$

Although we are suppressing the dependence on the sample in our notation, note that for each $A \subseteq S$, $P_n(A)$ is a *statistic* that gives the proportion of sample values in A .

21. Show that for fixed A , $P_n(A)$ is the sample mean from a random sample of size n from the distribution of the indicator random variable $\mathbf{1}(X \in A)$. Thus, conclude that

- $\mathbb{E}(P_n(A)) = \mathbb{P}(X \in A)$
- $\text{var}(P_n(A)) = \frac{1}{n} \mathbb{P}(X \in A) (1 - \mathbb{P}(X \in A))$
- $P_n(A) \rightarrow \mathbb{P}(X \in A)$ as $n \rightarrow \infty$ with probability 1.

This special case of the law of large numbers is central to the very concept of probability.

22. Show that for a fixed sample, P_n satisfies the axioms of a **probability measure**.

The empirical distribution of X , based on the random sample, is a random, discrete distribution, concentrated at the distinct sample values $\{X_1, X_2, \dots, X_n\}$. Indeed, it places probability mass $\frac{1}{n}$ at X_i for each $i \in \{1, 2, \dots, n\}$, so that if the sample values are distinct, the empirical distribution is uniform on these sample values.

Several applets in this project are simulations of random experiments with events of interest. When you run the experiment, you are performing independent replications of the experiment. In most cases, the applet displays the relative frequency of the event and its complement, both graphically in blue, and numerically in a

table. When you run the experiment, the relative frequencies are shown graphically in red and also numerically.

23. In the simulation of **Buffon's coin experiment**, the event of interest is that the coin crosses a crack. Run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the relative frequency of the event to the true probability.

24. In the simulation of **Bertrand's experiment**, the event of interest is that a “random chord” on a circle will be longer than the length of a side of the inscribed equilateral triangle. Run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the relative frequency of the event to the true probability.

The Empirical Distribution Function

Suppose now that X is a real-valued random variable for a basic experiment. Recall that the **distribution function** of X is the function F given by

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

Now suppose that we repeat the basic experiment n times to form a random sample (X_1, X_2, \dots, X_n) of size n from the distribution of X . It is natural to define the **empirical distribution function** for this sample by

$$F_n(x) = \frac{1}{n} \#(\{i \in \{1, 2, \dots, n\} : X_i \leq x\}), \quad x \in \mathbb{R}$$

Although we are suppressing the dependence on the sample in our notation, note that for each $x \in \mathbb{R}$, $F_n(x)$ is a *statistic* that gives the proportion of the sample variables that are less than or equal to x .

25. Show that F_n is the distribution function of the empirical distribution of X , based on the random sample (X_1, X_2, \dots, X_n) . In particular,

- F_n increases from 0 to 1.
- F_n is a step function with jumps at the distinct sample values $\{X_1, X_2, \dots, X_n\}$.

26. Show that for each x , $F_n(x)$ is the sample mean from a random sample of size n from the distribution of the indicator variable $\mathbf{1}(X \leq x)$. Thus, conclude that

- $\mathbb{E}(F_n(x)) = F(x)$
- $\text{var}(F_n(x)) = \frac{1}{n} F(x)(1 - F(x))$
- $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$ with probability 1.

Empirical Density for a Discrete Variable

Suppose now that X is a random variable for a basic experiment with a **discrete distribution** on a countable set S . Note that no assumptions are imposed on S , other than countability, so that in particular, X may be vector-valued. Recall that the **probability density function** of X is the function f given by

$$f(x) = \mathbb{P}(X = x), \quad x \in S$$

Now suppose that we repeat the basic experiment n times to form a random sample (X_1, X_2, \dots, X_n) of size n from the distribution of X . It is natural to define the **relative frequency function** or **empirical density function** of X for this sample as follows:

$$f_n(x) = \frac{1}{n} \#(\{i \in \{1, 2, \dots, n\} : X_i = x\}), \quad x \in S$$

Although we are suppressing the dependence on the sample in our notation, note that for each $x \in S$, $f_n(x)$ is a *statistic* that gives the proportion of the sample variables that have the value x .

27. Show that f_n is the probability density function of the empirical distribution of X , based on the sample (X_1, X_2, \dots, X_n) . In particular,

- $f_n(x) \geq 0$ for $x \in S$
- $\sum_{x \in S} f_n(x) = 1$

28. Show that for each x , $f_n(x)$, is the sample mean from a random sample of size n from the distribution of the indicator variable $\mathbf{1}(X = x)$. Thus, conclude that

- $\mathbb{E}(f_n(x)) = f(x)$
- $\text{var}(f_n(x)) = \frac{1}{n} f(x)(1 - f(x))$
- $f_n(x) \rightarrow f(x)$ as $n \rightarrow \infty$ with probability 1.

29. Show that if X is real-valued, then the sample mean of the random sample (X_1, X_2, \dots, X_n) is the mean computed relative to the empirical density function. That is,

$$\frac{1}{n} \sum_{i=1}^n X_i = \sum_{x \in S} x f_n(x)$$

Many of the applets in this project are simulations of experiments which result in discrete variables. When you run the simulation, you are performing independent replications of the experiment. In most cases, the applet displays the true density function numerically in a table and visually as a blue bar graph. When you run the simulation, the relative frequency function is also shown numerically in the table and visually as a red bar graph.

30. In the **poker experiment**, the random variable is the type of hand. Run the simulation 1000 times updating every 10 runs and note the apparent convergence of the empirical density function to the true

density function.

31. In the simulation of 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 empirical density function to the true density function.
32. 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 empirical density function to the true density function.

Empirical Density for a Continuous Variable

Recall again that the standard k -dimensional measure on \mathbb{R}^k is given by

$$\lambda_k(A) = \int_A 1 dx, \quad A \subseteq \mathbb{R}^k$$

In particular λ_1 is the length measure on \mathbb{R} , λ_2 is the area measure on \mathbb{R}^2 , and λ_3 is the volume measure on \mathbb{R}^3 .

Suppose now that X is a random variable for a basic experiment, with a **continuous distribution** on $S \subseteq \mathbb{R}^k$, and that X has **probability density function** f . Technically, f is the probability density function with respect to λ_k . Thus, by definition,

$$\mathbb{P}(X \in A) = \int_A f(x) dx, \quad A \subseteq S$$

Again we repeat the basic experiment n times to form a random sample (X_1, X_2, \dots, X_n) of size n from the distribution of X . Suppose now that $\{A_j : j \in J\}$ is a partition of S into a countable number of subsets. As before, we can define the empirical probability of A_j , based on the sample, by

$$P_n(A_j) = \frac{1}{n} \#(\{i \in \{1, 2, \dots, n\} : X_i \in A_j\})$$

We then define the **empirical density function** as follows:

$$f_n(x) = \frac{P_n(A_j)}{\lambda_k(A_j)}, \quad x \in A_j, j \in J$$

Clearly the empirical density function depends on the partition, as well as the sample, but we suppress this to keep the notation from becoming completely unwieldy. Of course, for each x , $f_n(x)$ is a random variable (in fact, a statistic), but by the very definition of density, if the partition is sufficiently fine (so that $\lambda_k(A_j)$ is small for each j), and if the sample size n is sufficiently large, then by the law of large numbers,

$$f_n(x) \approx f(x), \quad x \in S$$

33. Show that the empirical density function f_n satisfies the mathematical properties of a probability density function for a continuous distribution on S :

- $f_n(x) \geq 0$ for $x \in S$
- $\int_S f_n(x) dx = 1$

34. In fact, show that the empirical density function f_n corresponds to the distribution for which $P_n(A_j)$ is uniformly distributed over A_j for each $j \in J$

Many of the applets in this project are simulations of experiments which result in variables with continuous distributions. When you run the simulation, you are performing independent replications of the experiment. In most cases, the applet displays the true density visually as a blue graph. When you run the simulation, an empirical density function is also shown visually as a red bar graph.

35. In the simulation of the **gamma experiment**, vary the parameters and note the shape and location of the probability density function. For selected values of the parameters, run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the empirical density function to the true density function.

36. In the simulation of the **random variable experiment**, select the **normal distribution**. Run the experiment 1000 times with an update frequency of 10, and note the apparent convergence of the empirical density function to the true density function.

Exploratory Data Analysis

Many of the concepts discussed above are frequently used in **exploratory data analysis**. Specifically, suppose that x is a variable for a population (generally vector valued), and that (x_1, x_2, \dots, x_n) are the observed data from a sample of size n , corresponding to this variable. For example, x might be encode the color counts and net weight for a bag of M&Ms. Now let $\{A_j : j \in J\}$ be a partition of the data set, where J is a finite index set. The sets in the partition are generally known as **classes**. Just as above, we define the **frequency** of A_j by

$$\#(\{i \in \{1, 2, \dots, n\} : x_i \in A_j\})$$

We define the **relative frequency** of A_j by

$$\frac{1}{n} \#(\{i \in \{1, 2, \dots, n\} : x_i \in A_j\})$$

Finally, if x is a continuous variable, taking values in \mathbb{R}^k , we define the **density** of A_j by

$$\frac{1}{n \lambda_k(A_j)} \#(\{i \in \{1, 2, \dots, n\} : x_i \in A_j\})$$

The mapping that assigns frequencies to classes is known as a **frequency distribution** for the data set. The mapping that assigns relative frequencies to classes is known as a **relative frequency distribution** for the data set. Finally, in the case of a continuous variable, the mapping that assigns densities to classes is known as a **density distribution** for the data set. When $k = 1$ or $k = 2$, the bar graph of any of these distributions is known as a **histogram**.

The whole purpose of constructing and graphing one of these empirical distributions is to summarize and display the data in a meaningful way. Thus, there are some general guidelines in choosing the classes:

1. The number of classes should be moderate.
2. If possible, the classes should have the same size.

37. In the **interactive histogram**, click on the x -axis at various points to generate a data set with 20 values. Vary the class width over the five values from 0.1 to 5.0 and then back again. For each choice of class width, switch between the frequency histogram and the relative frequency histogram. Note how the shape of the histogram changes as you perform these operations.

It is important to realize that frequency data is inevitable for a continuous variable. For example, suppose that our variable represents the weight of a bag of M&M's (in grams) and that our measuring device (a scale) is accurate to 0.01 grams. If we measure the weight of a bag as 50.32, then we are really saying that the weight is in the interval $[50.315, 50.324)$ (or perhaps some other interval, depending on how the measuring device works). Similarly, when two bags have the same *measured* weight, the apparent equality of the weights is really just an artifact of the imprecision of the measuring device; actually the two bags almost certainly do *not* have the exact same weight. Thus, two bags with the same measured weight really give us a frequency count of 2 for a certain interval.

Again, there is a tradeoff between the *number* of classes and the *size* of the classes; these determine the **resolution** of the empirical distribution. At one extreme, when the class size is smaller than the accuracy of the recorded data, each class contains a single datum or no datum. In this case, there is no loss of information and we can recover the original data set from the frequency distribution (except for the *order* in which the data values were obtained). On the other hand, it can be hard to discern the shape of the data when we have many classes with small frequency. At the other extreme is a frequency distribution with one class that contains all of the possible values of the data set. In this case, all information is lost, except the number of the values in the data set. Between these two extreme cases, an empirical distribution gives us partial information, but not complete information. These intermediate cases can organize the data in a useful way.

38. In the **interactive histogram**, set the class width to 0.1. click on the x -axis to generate a data set with 10 distinct values and 20 values total.

- a. From the frequency distribution, explicitly write down the 20 values in the data set.
- b. Now increase the class width to 0.2, 0.5, 1.0, and 5.0. Note how the histogram loses resolution; that is, how the frequency distribution loses information about the original data set.

39. In **Michelson's data**, construct a frequency distribution for the velocity of light variable. Use 10 classes of equal width. Draw the histogram and describe the shape of the distribution.

40. In **Cavendish's data**, construct a relative frequency distribution for the density of the earth variable . Use 5 classes of equal width. Draw the histogram and describe the shape of the distribution.
41. In the **M&M data**, construct a frequency distribution and histogram for the total count variable and for the net weight variable.
42. In the **Cicada data**, construct a density distribution and histogram for the body weight variable for the cases given below . Note any differences.
- All cases
 - Each species individually
 - Male and female individually.
43. In the **interactive histogram**, set the class width in the to 0.1 and click on the axis to generate a distribution of the given type with 30 points. Now increase the class width to each of the other four values and describe the type of distribution.
- 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.

[Virtual Laboratories](#) > [6. Random Samples](#) > 1 **2** 3 4 5 6 7

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Key words](#) | [Feedback](#) | ©