6.2: The Sample Mean

Basic Theory

Recall the basic model of statistics: we have a population of objects of interest, and we have various measurements (variables) that we make on the objects. We select objects from the population and record the variables for the objects in the sample; these become our data. Our first discussion is from a purely descriptive point of view. That is, we do not assume that the data are generated by an underlying probability distribution. However, recall that the data themselves define a probability distribution.

Definition and Basic Properties

Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n from a real-valued variable. The *sample mean* is simply the arithmetic average of the sample values:

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{6.2.1}$$

If we want to emphasize the dependence of the mean on the data, we write $m(\boldsymbol{x})$ instead of just m. Note that m has the same physical units as the underlying variable. For example, if we have a sample of weights of cicadas, in grams, then m is in grams also. The sample mean is frequently used as a *measure of center* of the data. Indeed, if each x_i is the location of a *point mass*, then m is the *center of mass* as defined in physics. In fact, a simple graphical display of the data is the *dotplot*: on a number line, a dot is placed at x_i for each i. If values are repeated, the dots are stacked vertically. The sample mean m is the balance point of the dotplot. The image below shows a dot plot with the mean as the balance point.

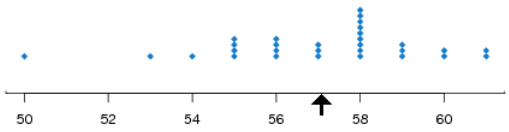

Figure 6.2.1: A dotplot

The standard notation for the sample mean corresponding to the data $\boldsymbol{x}$ is $\bar{x}$. We break with tradition and do not use the bar notation in this text, because it's clunky and because it's inconsistent with the notation for other statistics such as the sample variance, sample standard deviation, and sample covariance. However, you should be aware of the standard notation, since you will undoubtedly see it in other sources.

The following exercises establish a few simple properties of the sample mean. Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ are samples of size n from real-valued population variables and that c is a constant. In vector notation, recall that $\boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$ and $c\boldsymbol{x} = (cx_1, cx_2, \ldots, cx_n)$.

Computing the sample mean is a *linear operation*.

1. $m(\boldsymbol{x} + \boldsymbol{y}) = m(\boldsymbol{x}) + m(\boldsymbol{y})$
2. $m(c\,\boldsymbol{x}) = c\,m(\boldsymbol{x})$

Proof

1. $$m(\boldsymbol{x} + \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i + y_i) = \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{n} \sum_{i=1}^{n} y_i = m(\boldsymbol{x}) + m(\boldsymbol{y}) \tag{6.2.2}$$
2. $$m(c\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} cx_i = c \frac{1}{n} \sum_{i=1}^{n} x_i = cm(\boldsymbol{x}) \tag{6.2.3}$$

The sample mean preserves order.

1. If $x_i \geq 0$ for each i then $m(\boldsymbol{x}) \geq 0$.
2. If $x_i \geq 0$ for each i and $x_j > 0$ for some j them $m(\boldsymbol{x}) > 0$

3. If $x_i \le y_i$ for each i then $m(\boldsymbol{x}) \le m(\boldsymbol{y})$
4. If $x_i \le y_i$ for each i and $x_j < y_j$ for some j then $m(\boldsymbol{x}) < m(\boldsymbol{y})$

Proof

Parts (a) and (b) are obvious from the definition. Part (c) follows from part (a) and the linearity of expected value. Specifically, if $\boldsymbol{x} \le \boldsymbol{y}$ (in the product ordering), then $\boldsymbol{y} - \boldsymbol{x} \ge 0$. Hence by (a), $m(\boldsymbol{y} - \boldsymbol{x}) \ge 0$. But $m(\boldsymbol{y} - \boldsymbol{x}) = m(\boldsymbol{y}) - m(\boldsymbol{x})$. Hence $m(\boldsymbol{y}) \ge m(\boldsymbol{x})$. Similarly, (d) follows from (b) and the linearity of expected value.

Trivially, the mean of a constant sample is simply the constant. .

If $\boldsymbol{c} = (c, c, \ldots, c)$ is a constant sample then $m(\boldsymbol{c}) = c$.

Proof

Note that

$$m(\boldsymbol{c}) = \frac{1}{n} \sum_{i=1}^{n} c_i = \frac{nc}{n} = c \tag{6.2.4}$$

As a special case of these results, suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n corresponding to a real variable x, and that a and b are constants. Then the sample corresponding to the variable $y = a + bx$, in our vector notation, is $\boldsymbol{a} + b\boldsymbol{x}$. The sample means are related in precisely the same way, that is, $m(\boldsymbol{a} + b\boldsymbol{x}) = a + bm(\boldsymbol{x})$. Linear transformations of this type, when $b > 0$, arise frequently when physical units are changed. In this case, the transformation is often called a *location-scale* transformation; a is the location parameter and b is the scale parameter. For example, if x is the length of an object in inches, then $y = 2.54x$ is the length of the object in centimeters. If x is the temperature of an object in degrees Fahrenheit, then $y = \frac{5}{9}(x - 32)$ is the temperature of the object in degree Celsius.

Sample means are ubiquitous in statistics. In the next few paragraphs we will consider a number of special statistics that are based on sample means.

The Empirical Distribution

Suppose now that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n from a general variable taking values in a set S. For $A \subseteq S$, the *frequency* of A corresponding to $\boldsymbol{x}$ is the number of data values that are in A:

$$n(A) = \#\{i \in \{1, 2, \ldots, n\} : x_i \in A\} = \sum_{i=1}^{n} \mathbf{1}(x_i \in A) \tag{6.2.5}$$

The *relative frequency* of A corresponding to $\boldsymbol{x}$ is the proportion of data values that are in A:

$$p(A) = \frac{n(A)}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \in A) \tag{6.2.6}$$

Note that for fixed A, $p(A)$ is itself a sample mean, corresponding to the data $\{\mathbf{1}(x_i \in A) : i \in \{1, 2, \ldots, n\}\}$. This fact bears repeating: *every sample proportion is a sample mean*, corresponding to an indicator variable. In the picture below, the red dots represent the data, so $p(A) = 4/15$.

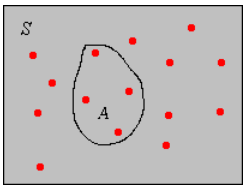

Figure 6.2.2: The empirical probability of A

p is a probability measure on S.

1. $p(A) \ge 0$ for every $A \subseteq S$
2. $p(S) = 1$

3. If $\{A_j : j \in J\}$ is a countable collection of pairwise disjont subsets of S then $p\left(\bigcup_{j \in J} A_j\right) = \sum_{j \in J} p(A_j)$

Proof

Parts (a) and (b) are obvious. For part (c) note that since the sets are disjoint,

$$p\left(\bigcup_{i \in I} A_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(x_i \in \bigcup_{j \in J} A_j\right) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in J} \mathbf{1}(x_i \in A_j) \tag{6.2.7}$$

$$= \sum_{j \in J} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \in A_j) = \sum_{j \in J} p(A_j) \tag{6.2.8}$$

This probability measure is known as the *empirical probability distribution* associated with the data set $\boldsymbol{x}$. It is a discrete distribution that places probability $\frac{1}{n}$ at each point x_i. In fact this observation supplies a simpler proof of previous theorem. Thus, if the data values are distinct, the empirical distribution is the discrete uniform distribution on $\{x_1, x_2, \ldots, x_n\}$. More generally, if $x \in S$ occurs k times in the data then the empirical distribution assigns probability k/n to x.

If the underlying variable is real-valued, then clearly the sample mean is simply the mean of the empirical distribution. It follows that the sample mean satisfies *all* properties of expected value, not just the linear properties and increasing properties given above. These properties are just the most important ones, and so were repeated for emphasis.

Empirical Density

Suppose now that the population variable x takes values in a set $S \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}_+$. Recall that the *standard measure* on $\mathbb{R}^d$ is given by

$$\lambda_d(A) = \int_A 1 \, dx, \quad A \subseteq \mathbb{R}^d \tag{6.2.9}$$

In particular $\lambda_1(A)$ is the length of A, for $A \subseteq \mathbb{R}$; $\lambda_2(A)$ is the area of A, for $A \subseteq \mathbb{R}^2$; and $\lambda_3(A)$ is the volume of A, for $A \subseteq \mathbb{R}^3$. Suppose that x is a continuous variable in the sense that $\lambda_d(S) > 0$. Typically, S is an interval if $d = 1$ and a Cartesian product of intervals if $d > 1$. Now for $A \subseteq S$ with $\lambda_d(A) > 0$, the *empirical density* of A corresponding to $\boldsymbol{x}$ is

$$D(A) = \frac{p(A)}{\lambda_d(A)} = \frac{1}{n \, \lambda_d(A)} \sum_{i=1}^{n} \mathbf{1}(x_i \in A) \tag{6.2.10}$$

Thus, the empirical density of A is the proportion of data values in A, divided by the size of A. In the picture below (corresponding to $d = 2$), if A has area 5, say, then $D(A) = 4/75$.

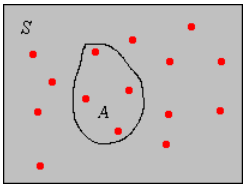

Figure 6.2.3: The empirical density of A

The Empirical Distribution Function

Suppose again that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n from a real-valued variable. For $x \in \mathbb{R}$, let $F(x)$ denote the relative frequency (empirical probability) of $(-\infty, x]$ corresponding to the data set $\boldsymbol{x}$. Thus, for each $x \in \mathbb{R}$, $F(x)$ is the sample mean of the data $\{\mathbf{1}(x_i \leq x) : i \in \{1, 2, \ldots, n\}\}$:

$$F(x) = p\left((-\infty, x]\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x) \tag{6.2.11}$$

F is a distribution function.

1. F increases from 0 to 1.
2. F is a step function with jumps at the distinct sample values $\{x_1, x_2, \ldots, x_n\}$.

Proof

Suppose that $(y_1, y_2, \ldots, y_k)$ are the distinct values of the data, ordered from smallest to largest, and that y_j occurs n_j times in the data. Then $F(x) = 0$ for $x < y_1$, $F(x) = n_1/n$ for $y_1 \leq x < y_2$, $F(x) = (n_1 + n_2)/n$ for $y_2 \leq x < y_3$, and so forth.

Appropriately enough, F is called the *empirical distribution function* associated with $\boldsymbol{x}$ and is simply the distribution function of the empirical distribution corresponding to $\boldsymbol{x}$. If we know the sample size n and the empirical distribution function F, we can recover the data, except for the order of the observations. The distinct values of the data are the places where F jumps, and the number of data values at such a point is the size of the jump, times the sample size n.

The Empirical Discrete Density Function

Suppose now that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n from a discrete variable that takes values in a countable set S. For $x \in S$, let $f(x)$ be the relative frequency (empirical probability) of x corresponding to the data set $\boldsymbol{x}$. Thus, for each $x \in S$, $f(x)$ is the sample mean of the data $\{\mathbf{1}(x_i = x) : i \in \{1, 2, \ldots, n\}\}$:

$$f(x) = p(\{x\}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i = x) \tag{6.2.12}$$

In the picture below, the dots are the possible values of the underlying variable. The red dots represent the data, and the numbers indicate repeated values. The blue dots are possible values of the the variable that did not happen to occur in the data. So, the sample size is 12, and for the value x that occurs 3 times, we have $f(x) = 3/12$.

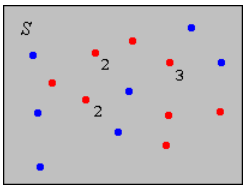

Figure 6.2.4: The discrete probability density function

f is a discrete probabiltiy density function:

1. $f(x) \geq 0$ for $x \in S$
2. $\sum_{x \in S} f(x) = 1$

Proof

Part (a) is obvious. For part (b), note that

$$\sum_{x \in S} f(x) = \sum_{x \in S} p(\{x\}) = 1 \tag{6.2.13}$$

Appropriately enough, f is called the *empirical probability density function* or the *relative frequency function* associated with $\boldsymbol{x}$, and is simply the probabiltiy density function of the empirical distribution corresponding to $\boldsymbol{x}$. If we know the empirical PDF f and the sample size n, then we can recover the data set, except for the order of the observations.

If the underlying population variable is real-valued, then the sample mean is the expected value computed relative to the empirical density function. That is,

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{x \in S} x \, f(x) \tag{6.2.14}$$

Proof

Note that

$$\sum_{x \in S} x f(x) = \sum_{x \in S} x \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i = x) = \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in S} x \mathbf{1}(x_i = x) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{6.2.15}$$

As we noted earlier, if the population variable is real-valued then the sample mean is the mean of the empirical distribution.

The Empirical Continuous Density Function

Suppose now that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n from a continuous variable that takes values in a set $S \subseteq \mathbb{R}^d$. Let $\mathscr{A} = \{A_j : j \in J\}$ be a partition of S into a countable number of subsets, each of positive, finite measure. Recall that the word *partition* means that the subsets are pairwise disjoint and their union is S. Let f be the function on S defined by the rule that $f(x)$ is the empricial density of A_j, corresponding to the data set $\boldsymbol{x}$, for each $x \in A_j$. Thus, f is constant on each of the partition sets:

$$f(x) = D(A_j) = \frac{p(A_j)}{\lambda_d(A_j)} = \frac{1}{n \lambda_d(A_j)} \sum_{i=1}^{n} \mathbf{1}(x_i \in A_j), \quad x \in A_j \tag{6.2.16}$$

f is a continuous probabiltiy density function.

1. $f(x) \geq 0$ for $x \in S$
2. $\int_S f(x) \, dx = 1$

Proof

Part (a) is obvious. For part (b) note that since f is constant on A_j for each $j \in J$ we have

$$\int_S f(x) \, dx = \sum_{j \in J} \int_{A_j} f(x) \, dx = \sum_{j \in J} \lambda_k(A_j) \frac{p(A_j)}{\lambda_k(A_j)} = \sum_{j \in J} p(A_j) = 1 \tag{6.2.17}$$

The function f is called the *empirical probability density function* associated with the data $\boldsymbol{x}$ and the partition $\mathscr{A}$. For the probability distribution defined by f, the empirical probability $p(A_j)$ is uniformly distributed over A_j for each $j \in J$. In the picture below, the red dots represent the data and the black lines define a partition of S into 9 rectangles. For the partition set A in the upper right, the empirical distribution would distribute probability $3/15 = 1/5$ uniformly over A. If the area of A is, say, 4, then $f(x) = 1/20$ for $x \in A$.

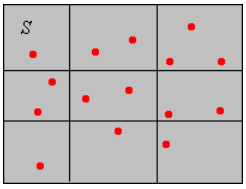

Figure 6.2.5: Empirical probability density function

Unlike the discrete case, we cannot recover the data from the empirical PDF. If we know the sample size, then of course we can determine the number of data points in A_j for each j, but not the precise location of these points in A_j. For this reason, the mean of the empirical PDF is not in general the same as the sample mean when the underlying variable is real-valued.

Histograms

Our next discussion is closely related to the previous one. Suppose again that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sample of size n from a variable that takes values in a set S and that $\mathscr{A} = (A_1, A_2, \ldots, A_k)$ is a partition of S into k subsets. The sets in the partition are sometimes known as *classes*. The underlying variable may be discrete or continuous.

- The mapping that assigns frequencies to classes is known as a *frequency distribution* for the data set and the given partition.
- The mapping that assigns relative frequencies to classes is known as a *relative frequency distribution* for the data set and the given partition.
- In the case of a continuous variable, the mapping that assigns densities to classes is known as a *density distribution* for the data set and the given partition.

In dimensions 1 or 2, the bar graph any of these distributions, is known as a *histogram*. The histogram of a frequency distribution and the histogram of the corresponding relative frequency distribution look the same, except for a change of scale on the vertical axis. If the classes all have the same size, the histogram of the corresponding density histogram also looks the same, again except for a change of scale on the vertical axis. If the underlying variable is real-valued, the classes are usually intervals (discrete or continuous) and the midpoints of these intervals are sometimes referred to as *class marks*.

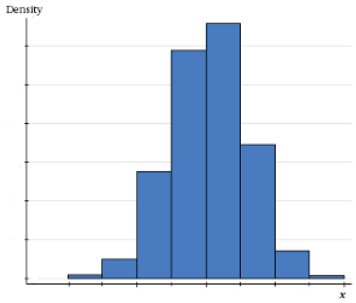

Figure 6.2.6: A density histogram

The whole purpose of constructing a partition and graphing one of these empirical distributions corresponding to the partition 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.

For highly skewed distributions, classes of different sizes are appropriate, to avoid numerous classes with very small frequencies. For a continuous variable with classes of different sizes, it is essential to use a density histogram, rather than a frequency or relative frequency histogram, otherwise the graphic is visually misleading, and in fact mathematically wrong.

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&Ms (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 trade-off between the *number* of classes and the *size* of the classes; these determine the *resolution* of the empirical distribution corresponding to the partition. 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.

Ogives

Suppose now the underlying variable is real-valued and that the set of possible values is partitioned into intervals $(A_1, A_2, \ldots, A_k)$, with the endpoints of the intervals ordered from smallest to largest. Let n_j denote the frequency of class A_j, so that $p_j = n_j / n$ is the relative frequency of class A_j. Let t_j denote the class mark (midpoint) of class A_j. The *cumulative frequency* of class A_j is $N_j = \sum_{i=1}^{j} n_i$ and the *cumulative relative frequency* of class A_j is $P_j = \sum_{i=1}^{j} p_i = N_j / n$. Note that the cumulative frequencies increase from n_1 to n and the cumulative relative frequencies increase from p_1 to 1.

- The mapping that assigns cumulative frequencies to classes is known as a *cumulative frequency distribution* for the data set and the given partition. The polygonal graph that connects the points (t_j, N_j) for $j \in \{1, 2, \ldots, k\}$ is the *cumulative frequency ogive*.
- The mapping that assigns cumulative relative frequencies to classes is known as a *cumulative relative frequency distribution* for the data set and the given partition. The polygonal graph that connects the points (t_j, P_j) for $j \in \{1, 2, \ldots, k\}$ is the *cumulative relative frequency ogive*.

Note that the relative frquency ogive is simply the graph of the distribution function corresponding to the probability distibution that places probability p_j at t_j for each j.

Approximating the Mean

In the setting of the last subsection, suppose that we do not have the actual data $\boldsymbol{x}$, but just the frequency distribution. An approximate value of the sample mean is

$$\frac{1}{n}\sum_{j=1}^{k} n_j t_j = \sum_{j=1}^{k} p_j t_j \tag{6.2.18}$$

This approximation is based on the hope that the mean of the data values in each class is close to the midpoint of that class. In fact, the expression on the right is the expected value of the distribution that places probability p_j on class mark t_j for each j.

Exercises

Basic Properties

Suppose that x is the temperature (in degrees Fahrenheit) for a certain type of electronic component after 10 hours of operation.

1. Classify x by type and level of measurement.
2. A sample of 30 components has mean 113°. Find the sample mean if the temperature is converted to degrees Celsius. The transformation is $y = \frac{5}{9}(x - 32)$.

Answer

1. continuous, interval
2. 45°

Suppose that x is the length (in inches) of a machined part in a manufacturing process.

1. Classify x by type and level of measurement.
2. A sample of 50 parts has mean 10.0. Find the sample mean if length is measured in centimeters. The transformation is $y = 2.54x$.

Answer

1. continuous, ratio
2. 25.4

Suppose that x is the number of brothers and y the number of sisters for a person in a certain population. Thus, $z = x + y$ is the number of siblings.

1. Classify the variables by type and level of measurement.
2. For a sample of 100 persons, $m(\boldsymbol{x}) = 0.8$ and $m(\boldsymbol{y}) = 1.2$. Find $m(\boldsymbol{z})$.

Answer

1. discrete, ratio
2. 2.0

Professor Moriarity has a class of 25 students in her section of Stat 101 at Enormous State University (ESU). The mean grade on the first midterm exam was 64 (out of a possible 100 points). Professor Moriarity thinks the grades are a bit low and is considering various transformations for increasing the grades. In each case below give the mean of the transformed grades, or state that there is not enough information.

1. Add 10 points to each grade, so the transformation is $y = x + 10$.
2. Multiply each grade by 1.2, so the transformation is $z = 1.2x$
3. Use the transformation $w = 10\sqrt{x}$. Note that this is a non-linear transformation that curves the grades greatly at the low end and very little at the high end. For example, a grade of 100 is still 100, but a grade of 36 is transformed to 60.

One of the students did not study at all, and received a 10 on the midterm. Professor Moriarity considers this score to be an outlier.

4. What would the mean be if this score is omitted?

Answer

1. 74
2. 76.8
3. Not enough information
4. 66.25

Computational Exercises

All statistical software packages will compute means and proportions, draw dotplots and histograms, and in general perform the numerical and graphical procedures discussed in this section. For real statistical experiments, particularly those with large data sets, the use of statistical software is essential. On the other hand, there is some value in performing the computations by hand, with small, artificial data sets, in order to master the concepts and definitions. In this subsection, do the computations and draw the graphs with minimal technological aids.

Suppose that x is the number of math courses completed by an ESU student. A sample of 10 ESU students gives the data $\boldsymbol{x} = (3, 1, 2, 0, 2, 4, 3, 2, 1, 2)$

1. Classify x by type and level of measurement.
2. Sketch the dotplot.
3. Compute the sample mean m from the definition and indicate its location on the dotplot.
4. Find the empirical density function f and sketch the graph.
5. Compute the sample mean m using f.
6. Find the empirical distribution function F and sketch the graph.

Answer

1. discrete, ratio
3. 2
4. $f(0) = 1/10$, $f(1) = 2/10$, $f(2) = 4/10$, $f(3) = 2/10$, $f(4) = 1/10$
5. 2
6. $F(x) = 0$ for $x < 0$, $F(x) = 1/10$ for $0 \le x < 1$, $F(x) = 3/10$ for $1 \le x < 2$, $F(x) = 7/10$ for $2 \le x < 3$, $F(x) = 9/10$ for $3 \le x < 4$, $F(x) = 1$ for $x \ge 4$

Suppose that a sample of size 12 from a discrete variable x has empirical density function given by $f(-2) = 1/12$, $f(-1) = 1/4$, $f(0) = 1/3$, $f(1) = 1/6$, $f(2) = 1/6$.

1. Sketch the graph of f.
2. Compute the sample mean m using f.
3. Find the empirical distribution function F
4. Give the sample values, ordered from smallest to largest.

Answer

2. $1/12$
3. $F(x) = 0$ for $x < -2$, $F(x) = 1/12$ for $-2 \le x < -1$, $F(x) = 1/3$ for $-1 \le x < 0$, $F(x) = 2/3$ for $0 \le x < 1$, $F(x) = 5/6$ for $1 \le x < 2$, $F(x) = 1$ for $x \ge 2$
4. $(-2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 2, 2)$

The following table gives a frequency distribution for the commuting distance to the math/stat building (in miles) for a sample of ESU students.

Class	Freq	Rel Freq	Density	Cum Freq	Cum Rel Freq	Midpoint
$(0, 2]$	6					
$(2, 6]$	16					
$(6, 10]$	18					
Total		**Density**	**Cum Freq**	**Cum Rel Freq**	**Midpoint**	

Class	Freq	Rel Freq	Density	Cum Freq	Cum Rel Freq	Midpoint
$(10, 20]$	10					
Total		**Density**	**Cum Freq**	**Cum Rel Freq**	**Midpoint**	

1. Complete the table
2. Sketch the density histogram
3. Sketch the cumulative relative frquency ogive.
4. Compute an approximation to the mean

Answer

1.

Class	Freq	Rel Freq	Density	Cum Freq	Cum Rel Freq	Midpoint
$(0, 2]$	6	0.12	0.06	6	0.12	1
$(2, 6]$	16	0.32	0.08	22	0.44	4
$(6, 10]$	18	0.36	0.09	40	0.80	8
$(10, 20])$	10	0.20	0.02	50	1	15
Total	50	1				

4. 7.28

App Exercises

In the interactive histogram, click on the x-axis at various points to generate a data set with at least 20 values. Vary the number of classes and switch between the frequency histogram and the relative frequency histogram. Note how the shape of the histogram changes as you perform these operations. Note in particular how the histogram loses resolution as you decrease the number of classes.

In the interactive histogram, click on the axis to generate a distribution of the given type with at least 30 points. Now vary the number of classes and note how the shape of the distribution changes.

1. A uniform distribution
2. A symmetric unimodal distribution
3. A unimodal distribution that is skewed right.
4. A unimodal distribution that is skewed left.
5. A symmetric bimodal distribution
6. A u-shaped distribution.

Data Analysis Exercises

Statistical software should be used for the problems in this subsection.

Consider the petal length and species variables in Fisher's iris data.

1. Classify the variables by type and level of measurement.
2. Compute the sample mean and plot a density histogram for petal length.
3. Compute the sample mean and plot a density histogram for petal length by species.

Answers

1. petal length: continuous, ratio. species: discrete, nominal
2. $m = 37.8$
3. $m(0) = 14.6,\ m(1) = 55.5,\ m(2) = 43.2$

Consider the erosion variable in the Challenger data set.

1. Classify the variable by type and level of measurement.
2. Compute the mean
3. Plot a density histogram with the classes $[0, 5)$, $[5, 40)$, $[40, 50)$, $[50, 60)$.

Answer

1. continuous, ratio
2. $m = 7.7$

Consider Michelson's velocity of light data.

1. Classify the variable by type and level of measurement.
2. Plot a density histogram.
3. Compute the sample mean.
4. Find the sample mean if the variable is converted to km/hr. The transformation is $y = x + 299\,000$

Answer

1. continuous, interval
3. $m = 852.4$
4. $m = 299\,852.4$

Consider Short's paralax of the sun data.

1. Classify the variable by type and level of measurement.
2. Plot a density histogram.
3. Compute the sample mean.
4. Find the sample mean if the variable is converted to degrees. There are 3600 seconds in a degree.
5. Find the sample mean if the variable is converted to radians. There are $\pi/180$ radians in a degree.

Answer

1. continuous, ratio
3. 8.616
4. 0.00239
5. 0.0000418

Consider Cavendish's density of the earth data.

1. Classify the variable by type and level of measurement.
2. Compute the sample mean.
3. Plot a density histogram.

Answer

1. continuous, ratio
2. $m = 5.448$

Consider the M&M data.

1. Classify the variables by type and level of measurement.
2. Compute the sample mean for each color count variable.
3. Compute the sample mean for the total number of candies, using the results from (b).
4. Plot a relative frequency histogram for the total number of candies.
5. Compute the sample mean and plot a density histogram for the net weight.

Answer

1. color counts: discrete ratio. net weight: continuous ratio.
2. $m(r) = 9.60$, $m(g) = 7.40$, $m(bl) = 7.23$, $m(o) = 6.63$, $m(y) = 13.77$, $m(br) = 12.47$
3. $m(n) = 57.10$
5. $m(w) = 49.215$

Consider the body weight, species, and gender variables in the Cicada data.

1. Classify the variables by type and level of measurement.
2. Compute the relative frequency function for species and plot the graph.
3. Compute the relative frequeny function for gender and plot the graph.
4. Compute the sample mean and plot a density histogram for body weight.
5. Compute the sample mean and plot a density histogrm for body weight by species.
6. Compute the sample mean and plot a density histogram for body weight by gender.

Answer

1. body weight: continuous, ratio. species: discrete, nominal. gender: discrete, nominal.
2. $f(0) = 0.423,\ f(1) = 0.519,\ f(2) = 0.058$
3. $f(0) = 0.567,\ f(1) = 0.433$
4. $m = 0.180$
5. $m(0) = 0.168,\ m(1) = 0.185,\ m(2) = 0.225$
6. $m(0) = 0.206,\ m(1) = 0.145$

Consider Pearson's height data.

1. Classify the variables by type and level of measurement.
2. Compute the sample mean and plot a density histogram for the height of the father.
3. Compute the sample mean and plot a density histogram for the height of the son.

Answer

1. continuous ratio
2. $m(f) = 67.69$
3. $m(s) = 68.68$

This page titled 6.2: The Sample Mean is shared under a CC BY 2.0 license and was authored, remixed, and/or curated by Kyle Siegrist (Random Services) via source content that was edited to the style and standards of the LibreTexts platform.