Median for continuous distribution

The expected value of a discrete random variable is the sum of all the values the variable can take times the probability of that value occurring. $$ {E} \left({X} \right)=\sum { {x}. {p} \left({x} \right)} $$ The expected value of random variable \(X\) is often written as \({E} \left({X} \right)\) or \({ \mu }\) or \({ \mu }\text{X}\). Example: Expected Return of a Discrete Random Variable Given the experiment of rolling a single die, calculate the expected value. Solution Using the result: $$ {E} \left({X} \right)=\sum { {x}.{p} \left({x} \right)} $$ We have: $$ {E} \left({X} \right) =1\times \left(\cfrac{1}{6}\right)+2\times \left(\cfrac{1}{6}\right)+3\times \left(\cfrac{1}{6}\right)+4\times \left(\cfrac{1}{6}\right)+5\times \left(\cfrac{1}{6}\right)+6\times \left(\cfrac{1}{6}\right)= 3.5 $$ Expected Value of a Continuous Random Variable If \(\alpha\) and \(\beta\) are constants, then: $$E(\alpha X)=a\bullet E(X)=a\bullet \sum{x\bullet P(x)}$$ And $$E(\alpha X+\beta)=\alpha E(X)+\beta$$ If \(X\) is a continuous random variable with a probability density function \({f}\left({x} \right)\), then the expected value (or mean) of \(X\) is given by $$ E\left(X \right) =\int _{ -\infty }^{ \infty }{ xf\left(x \right)dx } $$ Where \(f(x)\) is the probability density function of \(x\). Example: Expected Return of Continuous Random Variable Given the following probability density function of a continuous random variable: $$ f\left(x \right) =\begin{cases} \cfrac { x }{ 2 } , & 0< x < 2 \\ 0, & otherwise \end{cases} $$ Calculate \(E(X)\). $$ E\left(X \right) =\int _{ -\infty }^{ \infty }{ xf\left(x \right) dx } =\int _{ 0 }^{ 2 }{ x\bullet \frac { x }{ 2 } \bullet dx={ \left[\frac { { x }^{ 3 } }{ 6 } \right] }_{ x=0 }^{ x=2 } =\frac { 8 }{ 6 } =\frac { 4 }{ 3 } } $$ The Mode of a Discrete Random Variable The mode of a discrete random variable is the value that is most likely to occur. Example: Calculating the Mode of Discrete Random Variable Given the experiment of rolling two dice simultaneously, what is the mode of the probability distribution of the sum of the two dice? Solution As shown in the table below, the most likely value is 7 with a probability of \(\frac{6}{36}\) so the mode is 7. $$ \begin{array}{c|c} {\bf x} & {\bf p(x)} \\ \hline 2 & {1}/{36} \\ \hline 3 & {2}/{36} \\ \hline 4 & {3}/{36} \\ \hline 5 & {4}/{36} \\ \hline 6 & {5}/{36} \\ \hline 7 & {6}/{36} \\ \hline 8 & {5}/{36} \\ \hline 9 & {4}/{36} \\ \hline 10 & {3}/{36} \\ \hline 11 & {2}/{36} \\ \hline 12 & {1}/{36} \\ \end{array} $$ As shown in the table below, the most likely value is 7 with a probability of \(\cfrac{6}{36}\), so the mode = 7. The Mode of a Continuous Random Variable The mode of a continuous random variable is the value at which the probability density function, \(f(x)\), is at a maximum. It is a value that is most likely to lie within the same interval as the outcome. Consequently, often we will find the mode(s) of a continuous random variable by solving the equation: $$ {f}^{\prime} \left(\text{x} \right)=0 $$ Remember: For the case of continuous random variables, the probability of a specific value occurring is 0, \(P(X=k)=0\), and the mode is a specific value. Example: Calculating the Mode of a Distribution Given the following probability density function of a continuous random variable, find the mode of the distribution. $$ f\left(x \right) =\begin{cases} -{ x }^{ 2 }+2x-\cfrac { 1 }{ 6 } , & 0 < x < 2 \\ 0, & otherwise \end{cases} $$ To find the maximum of \(f(x)\), find the first derivative and set that value equal to zero as shown below: \({f}^{\prime} (x)= -2x+2=0 \) Then solve for \(x\). \(-2x = -2\) \(x = 1\) So, the mode of the distribution is 1. Median of a Discrete Random Distribution The median of a discrete distribution is the value of \(X\) for which \(P(X \le x)\) is greater than or equal to 0.5 and \(P(X \ge x)\) is greater than or equal to 0.5. Example: Calculating the Median of Discrete Random Distribution Given the following probability density function of a discrete random variable, calculate the median of the distribution: $$ {f}\left({x} \right) =\begin{cases} 0.2,{x}=1,4 & \\ 0.3,{x}=2,3, & \end{cases} $$ Solution The median of the distribution above is 2 because: $$ P(X \le 2) = P(X=1)+P(X=2) = .2+.3 = .5$$ and, $$P(X \ge 2) = P(X=2)+P(X=3)+P(X=4) = .3+.3+.2 = .8. $$ Median of a Continuous Distribution The median of a continuous distribution can be defined by the value \(c\) in the formula shown below: $$ \int _{ -\infty }^{ c }{ f\left(x \right)dx=0.5 } $$ That is, it is the value for which the area under the curve from negative infinity to \(c\) is equal to 0.5. The median is also referred to as the 50th percentile. Example: Calculating the Median of a Continuous Distribution Given the following probability density function of a continuous random variable, find the median of the distribution. $$ {f}\left({x} \right)=\begin{cases} -{x}^{ 2 }+2 {X}-\frac {1}{6}, & 0 < x < 2\\ 0, & \text{otherwise}\\ \end{cases}$$ Solution We need, $$ \begin{align*} & {P}\left({x}\le{{m}}\right)=0.5 \\ &\Rightarrow \int _{ -\infty }^{ {m} }{ \left(-{ {x} }^{2}+2{x}-\cfrac{1}{6}\right){dx}=0.5}\\ &=\left[-\cfrac{{ {x} }^{3}}{3}+{ {x} }^{2}-\cfrac{1}{6} { {x} }\right]_{ { {x} }=0}^{ { {x} }={ {m} }}=0.5 \\ &=-\cfrac{{ {m} }^{3}}{3}+{ {m} }^{2}-\cfrac{1}{6}\times {c}=0.5 \\ &\Rightarrow { {m} }=1 \end{align*} $$ pth Percentile of a Discrete Distribution In general, the pth percentile of a discrete distribution is the value of \(x\) for which \(P(X \le x)\) is greater than or equal to \(\frac {p}{100}\) and \(P(X \ge x)\) is greater than or equal to \(1-\frac {p}{100}\). Example: Calculating the 75th Percentile from Discrete Random Variables Given the following probability density function of a discrete random variable, calculate the 75th percentile of the distribution: $$ {f}\left({x} \right) =\begin{cases} 0.2,{x}=1,4 & \\ 0.3,{x}=3,4 & \end{cases} $$ Solution We need: $$ {P}\left({X} \le {{p}}_{0.75} \right) = .75 $$ So, $$ \begin{align*} {P}\left({X} \le {3} \right) & = {P}\left({X} = {1} \right) + {P}\left({X} = {2} \right) + {P}\left({X} = {3} \right)=0.2 + 0.3 + 0.2 \\ & = 0.7 < 0.75 \\ {P}\left({X} \le {4} \right) & = {P}\left({X} = {1} \right) + {P}\left({X} = {2} \right) + {P}\left({X} = {3} \right) + {P}\left({X} = {4} \right) \\ & = 0.2 +0.3 + 0.2 + 0.3 = 1.0 > 0.75 \\ {P}\left({X} \le {4} \right) & =({P}\left({X} = {4} \right)=0.2 < 0.25 = 1 – \left(\cfrac{75}{100} \right) \end{align*} $$ So, the 75th percentile of the distribution is 4. In general, the pth percentile of a continuous distribution can be defined as the value of \(C\) for which: $$ \int _{ -\infty }^{ \infty }{ {f}\left({x} \right) {dx} }=\cfrac{ {p} }{100} $$ Example: Calculating the 25th Percentile from Discrete Random Variables Given the following probability density function of a continuous random variable, find the 25th percentile: $$ {f}\left({x} \right)=\begin{cases} -\text{x}^2+2\text{x}-\frac {1}{6}, & 0 < \text{x} < 2\\ 0, & \text{elsewhere}\\ \end{cases}$$ Solution We need: $$ \begin{align*} &{P}\left({x}\le{{c}}\right)=0.25 \\ &\Rightarrow \int _{ 0 }^{ {c} }{ \left(-{ {x} }^{2}+2{x}-\cfrac{1}{6}\right){dx}=\cfrac{25}{100}=0.25}\\ &\int _{ 0 }^{ 2 }{ \left(-{ {x} }^{2}+2{x}-\cfrac{1}{6}\right) {dx}=0.25} \\ &\left[-\cfrac{ { {x} }^{3} }{3}+2 {x}-\cfrac{1}{6} {x} \right]_{{{x}}=0}^{{x}={c}}=0.25 \\ &-\cfrac{ { {c} }^{3} }{3}+{ {C} }^{2}-\cfrac{1}{6}\times{ {c} }=0.25 \\ &{ {C} }=0.69 \end{align*} $$ Learning Outcome Topic 2.c: Univariate Random Variables – Explain and calculate expected value, mode, median, percentile, and higher moments. The terms mean, median, mode, and range describe properties of statistical distributions. In statistics, a distribution is the set of all possible values for terms that represent defined events. The value of a term, when expressed as a variable, is called a random variable. There are two major types of statistical distributions. The first type contains discrete random variables. This means that every term has a precise, isolated numerical value. The second major type of distribution contains a continuous random variable. A continuous random variable is a random variable where the data can take infinitely many values. When a term can acquire any value within an unbroken interval or span, it is called a probability density function. IT professionals need to understand the definition of mean, median, mode and range to plan capacity and balance load, manage systems, perform maintenance and troubleshoot issues. Furthermore, understanding of statistical terms is important in the growing field of data science. Understanding the definition of mean, median, mode and range is important for IT professionals in data center management. Many relevant tasks require the administrator to calculate mean, median, mode or range, or often some combination, to show a statistically significant quantity, trend or deviation from the norm. Finding the mean, median, mode and range is only the start. The administrator then needs to apply this information to investigate root causes of a problem, accurately forecast future needs or set acceptable working parameters for IT systems. When working with a large data set, it can be useful to represent the entire data set with a single value that describes the "middle" or "average" value of the entire set. In statistics, that single value is called the central tendency and mean, median and mode are all ways to describe it. To find the mean, add up the values in the data set and then divide by the number of values that you added. To find the median, list the values of the data set in numerical order and identify which value appears in the middle of the list. To find the mode, identify which value in the data set occurs most often. Range, which is the difference between the largest and smallest value in the data set, describes how well the central tendency represents the data. If the range is large, the central tendency is not as representative of the data as it would be if the range was small. Mean The most common expression for the mean of a statistical distribution with a discrete random variable is the mathematical average of all the terms. To calculate it, add up the values of all the terms and then divide by the number of terms. The mean of a statistical distribution with a continuous random variable, also called the expected value, is obtained by integrating the product of the variable with its probability as defined by the distribution. The expected value is denoted by the lowercase Greek letter mu (μ). Median The median of a distribution with a discrete random variable depends on whether the number of terms in the distribution is even or odd. If the number of terms is odd, then the median is the value of the term in the middle. This is the value such that the number of terms having values greater than or equal to it is the same as the number of terms having values less than or equal to it. If the number of terms is even, then the median is the average of the two terms in the middle, such that the number of terms having values greater than or equal to it is the same as the number of terms having values less than or equal to it. The median of a distribution with a continuous random variable is the value m such that the probability is at least 1/2 (50%) that a randomly chosen point on the function will be less than or equal to m, and the probability is at least 1/2 that a randomly chosen point on the function will be greater than or equal to m. Mode The mode of a distribution with a discrete random variable is the value of the term that occurs the most often. It is not uncommon for a distribution with a discrete random variable to have more than one mode, especially if there are not many terms. This happens when two or more terms occur with equal frequency, and more often than any of the others. A distribution with two modes is called bimodal. A distribution with three modes is called trimodal. The mode of a distribution with a continuous random variable is the maximum value of the function. As with discrete distributions, there may be more than one mode. Range The range of a distribution with a discrete random variable is the difference between the maximum value and the minimum value. For a distribution with a continuous random variable, the range is the difference between the two extreme points on the distribution curve, where the value of the function falls to zero. For any value outside the range of a distribution, the value of the function is equal to 0. To calculate mean, add together all of the numbers in a set and then divide the sum by the total count of numbers. For example, in a data center rack, five servers consume 100 watts, 98 watts, 105 watts, 90 watts and 102 watts of power, respectively. The mean power use of that rack is calculated as (100 + 98 + 105 + 90 + 102 W)/5 servers = a calculated mean of 99 W per server. Intelligent power distribution units report the mean power utilization of the rack to systems management software. In the data center, means and medians are often tracked over time to spot trends, which inform capacity planning or power cost predictions.The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median. For the five servers in the rack, arrange the power consumption figures from lowest to highest: 90 W, 98 W, 100 W, 102 W and 105 W. The median power consumption of the rack is 100 W. If there is an even set of numbers, average the two middle numbers. For example, if the rack had a sixth server that used 110 W, the new number set would be 90 W, 98 W, 100 W, 102 W, 105 W and 110 W. Find the median by averaging the two middle numbers: (100 + 102)/2 = 101 W. The mode is the number that occurs most often within a set of numbers. For the server power consumption examples above, there is no mode because each element is different. But suppose the administrator measured the power consumption of an entire network operations center (NOC) and the set of numbers is 90 W, 104 W, 98 W, 98 W, 105 W, 92 W, 102 W, 100 W, 110 W, 98 W, 210 W and 115 W. The mode is 98 W since that power consumption measurement occurs most often amongst the 12 servers. Mode helps identify the most common or frequent occurrence of a characteristic. It is possible to have two modes (bimodal), three modes (trimodal) or more modes within larger sets of numbers. The range is the difference between the highest and lowest values within a set of numbers. To calculate range, subtract the smallest number from the largest number in the set. If a six-server rack includes 90 W, 98 W, 100 W, 102 W, 105 W and 110 W, the power consumption range is 110 W - 90 W = 20 W. Range shows how much the numbers in a set vary. Many IT systems operate within an acceptable range; a value in excess of that range might trigger a warning or alarm to IT staff. To find the variance in a data set, subtract each number from the mean, and then square the result. Find the average of these squared differences, and that is the variance in the group. In our original group of five servers, the mean was 99. The 100 W-server varies from the mean by 1 W, the 105 W-server by 6 W, and so on. The squares of each difference equal 1, 1, 36, 81 and 9. So to calculate the variance, add 1 + 1 + 36 + 81 + 9 and divide by 5. The variance is 25.6. Standard deviation denotes how far apart all the numbers are in a set. The standard deviation is calculated by finding the square root of the variance. In this example, the standard deviation is 5.1. Interquartile range, the middle fifty or midspread of a set of numbers, removes the outliers -- highest and lowest numbers in a set. If there is a large set of numbers, divide them evenly into lower and higher numbers. Then find the median of each of these groups. Find the interquartile range by subtracting the lower median from the higher median. If a rack of six servers' power wattage is arranged from lowest to highest: 90, 98, 100, 102, 105, 110, divide this set into low numbers (90, 98, 100) and high numbers (102, 105, 110). Find the median for each: 98 and 105. Subtract the lower median from the higher median: 105 watts - 98 W = 7 W, which is the interquartile range of these servers.