Department of Mathematics

The Statistics of a Function

Section 6.7

Dr. John Ehrke
Department of Mathematics

Spring 2013

The Idea of an Average

If five people weigh 155, 143, 180, 105, and 123 lb, their average (mean) weight is given by

$$\frac{155 + 143 + 180 + 105 + 123}{5} = 141.2 \text{ lb}.$$

This idea generalizes quite naturally to functions. Consider a function $f(x)$ that is continuous on $[a, b]$. Form a regular partition of $[a, b]$, $a = x_0 < x_1 < \ldots < x_n = b$ with $\Delta x = (b - a)/n$. Select a point $\bar{x}_k$ in each subinterval and compute $f(\bar{x}_k)$ for $k = 1, \ldots, n$. The values of $f(\bar{x}_k)$ is called a *sampling* of $f(x)$ on $[a, b]$. The average of these values is

$$\frac{f(\bar{x}_0) + f(\bar{x}_1) + f(\bar{x}_2) + \cdots + f(\bar{x}_n)}{n}.$$

Noting that $n = (b - a)/\Delta x$, we write the average of the n values at the Riemann sum

$$\frac{f(\bar{x}_0) + f(\bar{x}_1) + f(\bar{x}_2) + \cdots + f(\bar{x}_n)}{(b - a)/\Delta x} = \frac{1}{b - a} \sum_{k=0}^{n} f(\bar{x}_k) \Delta x.$$

The Average Value Integral

Now suppose we increase our partition points n, taking more and more samples of f, while Δx decreases to zero. The limit of this sum is a definite integral that gives the average value of f on $[a, b]$. In symbols,

$$\mu = \bar{f}(x) = \frac{1}{b-a} \int_a^b f(x)\, dx.$$

The average value of a function f on an interval $[a, b]$ has a clear geometric meaning. To see this, multiply both sides of the formula above by $(b - a)$. This says,

$$(b-a)\mu = \int_a^b f(x)\, dx.$$

What does this mean?

> *The average value is the height of the rectangle with base = $(b - a)$ on the interval $[a, b]$ that has the same area as the region bounded by the graph of f on the same interval.*

Average Elevation Example

Example

A hiking trail has an elevation given by

$$f(x) = 60x^3 - 650x^2 + 1200x + 4500$$

where f is measured in feet above sea level and x represents the horizontal trail distance in miles ($0 \leq x \leq 5$). What is the average elevation of the trail?

Average Elevation Example

Example

A hiking trail has an elevation given by

$$f(x) = 60x^3 - 650x^2 + 1200x + 4500$$

where f is measured in feet above sea level and x represents the horizontal trail distance in miles ($0 \leq x \leq 5$). What is the average elevation of the trail?

Solution: The trail ranges between elevation of about 2000 and 5000 feet. If we let the endpoints of the trail correspond to the horizontal distances $a = 0$ and $b = 5$ miles, the average elevation of the trail is

$$\begin{aligned}
\mu &= \frac{1}{5}\int_0^5 60x^3 - 650x^2 + 1200x + 4500\, dx \\
&= \frac{1}{5}\left(15x^4 - \frac{650}{3}x^3 + 600x^2 + 4500x\right)_0^5 \\
&= 3958.33 \text{ feet}
\end{aligned}$$

Mean Value Theorem for Integrals

The average value of a function brings us close to an important theoretical result.

Theorem (Mean Value Theorem for Integrals)

Let f be continuous on the interval $[a, b]$. There exists a point c in $[a, b]$ such that

$$f(c) = \mu = \frac{1}{b-a} \int_a^b f(x)\, dx.$$

Mean Value Theorem for Integrals

The average value of a function brings us close to an important theoretical result.

Theorem (Mean Value Theorem for Integrals)

Let f be continuous on the interval $[a, b]$. There exists a point c in $[a, b]$ such that

$$f(c) = \mu = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

Proof: Because f is continuous on a closed interval $[a, b]$, then it attains a minimum value y_{min} and a maximum value y_{max} on $[a, b]$. This is the result of the *Extreme Value Theorem* from Calculus I. Also, note that

$$(b-a)y_{min} \leq \int_a^b f(x)\,dx \leq (b-a)y_{max}.$$

Draw a picture to see this!! Dividing through these inequalities by $(b-a)$, we have

$$y_{min} \leq \frac{1}{b-a}\int_a^b f(x)\,dx \leq y_{max}.$$

By the intermediate value theorem, $f(x)$ takes on every value between y_{min} and y_{max} and so in particular takes on the value $\bar{f}$ for some $c \in [a, b]$ as claimed.

Variance of a Function

The variance measurement in statistics is a measure of the dispersion or spread of the values of the function about the mean value of the function $\bar{f}(x) = \mu$. If the number of subintervals is sufficiently large, then we could sample the interval at each of the $n+1$ end points, $f(x_0), f(x_1), f(x_2), \ldots, f(x_n)$ and using the variance formula from statistics we could calculate:

$$\text{Variance of } f(x) \approx \frac{1}{n}\sum_{k=0}^{n} f(x_k) = \frac{1}{b-a}\sum_{k=0}^{n} [f(x_i) - \mu]^2 \ \Delta x.$$

This is a Riemann sum and so taking the limit at $n \to \infty$ or as $\Delta x \to 0$, we find

$$\begin{aligned}
\text{Variance of } f(x) &= \frac{1}{b-a}\int_a^b [f(x) - \mu]^2 \ dx \\
&= \frac{1}{b-a}\int_a^b \left[f^2(x) - 2\mu f(x) + \mu^2\right] \ dx \\
&= \frac{1}{b-a}\int_a^b f^2(x) \ dx - 2\mu^2 + \frac{1}{b-a}\mu^2(b-a) \\
&= \frac{1}{b-a}\int_a^b f^2(x) \ dx - \mu^2.
\end{aligned}$$

Since the standard deviation σ is defined as the square root of the variance we now have a formula for the standard deviation of a function,

$$\sigma = \sqrt{\frac{1}{b-a}\int_a^b f^2(x) \ dx - \mu^2}.$$

Standard Deviation of a Function

Example

Consider the linear function $f(x) = \alpha x + \beta$ on the interval $[0, 1]$. Using the formulas derived in this lecture find the mean and standard deviation of this function.

Standard Deviation of a Function

Example

Consider the linear function $f(x) = \alpha x + \beta$ on the interval $[0, 1]$. Using the formulas derived in this lecture find the mean and standard deviation of this function.

Solution: The mean of the function is

$$\mu = \frac{1}{1-0}\int_0^1 (\alpha x + \beta)\, dx = \left(\frac{\alpha x^2}{2} + \beta\right)_0^1 + \frac{\alpha}{2} + \beta = f\left(\frac{1}{2}\right).$$

Thus as we might have expected the mean of the function is equal to the value at the midpoint of the interval. Calculating the standard deviation we obtain

$$\sigma = \sqrt{\frac{1}{1-0}\int_0^1 (\alpha x + \beta)^2\, dx - \left(\frac{\alpha}{2} + \beta\right)^2} = \frac{\alpha}{\sqrt{12}}.$$

Applications to Probability

Continuous Random Variable

A variable whose value is subject to change based on chance or randomness and whose values range over an entire interval with each value having an associated probability is called a continuous random variable.

Continuous random variables take on values that vary continuously within one or more (possibly infinite) intervals. As a result there are an uncountably infinite number of individual outcomes, and each has a probability of 0. As a result, the probability distribution for many continuous random variables is defined using a probability density function.

Probability Density Function

Let x be a continuous random variable distributed over some interval $[a, b]$. A function $f(x)$ is said to be a probability density function for x if:

1. $f(x)$ is non-negative on $[a, b]$
2. $\displaystyle\int_a^b f(x)\,dx = 1$
3. For any subinterval $[c, d]$ of $[a, b]$ the probability that x lies in $[c, d]$ is given by

$$P(c \leq x \leq d) = \int_c^d f(x)\,dx.$$

Finding a pdf

Example

Find k such that $f(x) = kx^2$ is a probability density function over the interval $[2, 5]$. Then find $P(2.5 \leq x \leq 2.75)$ and interpret the answer in the context of the previous example.

Finding a pdf

Example

Find k such that $f(x) = kx^2$ is a probability density function over the interval $[2, 5]$. Then find $P(2.5 \leq x \leq 2.75)$ and interpret the answer in the context of the previous example.

Solution: To find the value of k we first know that $k > 0$ since $f(x) > 0$ on $[2, 5]$. Secondly we need

$$\int_2^5 kx^2 \, dx = 1$$

which means that

$$\int_2^5 kx^2 \, dx = \frac{kx^3}{3}\bigg|_2^5 = \frac{125k}{3} - \frac{8k}{3} = \frac{117k}{3} \Longrightarrow k = \frac{3}{117}.$$

So the pdf is $f(x) = \frac{3}{117}x^2$ and so the probability

$$P(2.5 \leq x \leq 2.75) = \int_{2.5}^{2.75} \frac{3}{117}x^2 \, dx \approx 0.04420405983$$

A Curious Example

Let us consider again throwing the dart at a number line in such a way that it lands in the interval $[a, b]$. This time we will assume it is equally likely for the dart to land anywhere in the interval. Suppose we threw the dart n times and kept track of the numbers it hits. The average of these numbers would look like,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} x_i \cdot \frac{1}{n}.$$

The expression

$$\sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \sum_{i=1}^{n} x_i \cdot \frac{1}{b-a} \cdot \frac{b-a}{n}.$$

Note, that $1/(b-a)$ is a probability density function for $[a, b]$ and $(b-a)/n = \Delta x$. This gives,

$$\bar{x} = \sum_{i=1}^{n} x_i \cdot f(x_i)\Delta x \to \int_a^b x \cdot f(x)\, dx$$

where $f(x)$ is the probability density function for the continuous random variable x.

Example

Example

Suppose that we have the probability density function, $f(x) = \frac{1}{4}x$ over the interval $[1, 3]$. Find and interpret the area under under the curve in the context of the dart problem. Using the method above compute the average landing spot for the dart.

Example

Example

Suppose that we have the probability density function, $f(x) = \frac{1}{4}x$ over the interval $[1, 3]$. Find and interpret the area under under the curve in the context of the dart problem. Using the method above compute the average landing spot for the dart.

Solution: This function, $f(x) = 0.25x$ gives more "weight" to the right hand side of the interval than to the left. Perhaps more points are awarded when the dart hits on the right. Then

$$\int_1^3 x \cdot f(x)\, dx = \int_1^3 x \cdot \frac{x}{4}\, dx \approx 2.17.$$

Suppose we continue to throw the dart and compute average. The more times we throw the dart, the closer we expect the averages to come to 2.17.

Expected Value

Let x be a continuous random variable over the interval $[a, b]$ with probability density function $f(x)$.

Definition

The *expected value* of x is defined as

$$E(x) = \int_a^b x \cdot f(x)\, dx.$$

The concept of expected value of a random variable generalizes to other functions of a random variable. Suppose that $y = g(x)$ is a function of the random variable x. The expected value of $g(x)$ is defined as

$$E(x) = \int_a^b g(x) \cdot f(x)\, dx.$$

The mean μ of a continuous random variable x is defined to $E(x)$. That is

$$\mu = E(x) = \int_a^b x \cdot f(x)\, dx.$$

Variance and Standard Deviation

Definition

The variance σ^2 of a continuous random variable x is defined as

$$\sigma^2 = E(x^2) - \mu^2 = E(x^2) - [E(x)]^2 = \int_a^b x^2 f(x)\,dx - \left[\int_a^b xf(x)\,dx\right]^2.$$

The standard deviation σ of a continuous random variable is defined as the square root of the variance.

Example

Given the probability density function $f(x) = \frac{1}{2}x$ over $[0, 2]$ find the mean, variance, and standard deviation.

Solution

Calculating $E(x)$, we have

$$E(x) = \int_0^2 x \cdot \frac{x}{2}\, dx = \frac{4}{3}.$$

Likewise for $E(x^2)$, we have

$$E(x^2) = \int_0^2 x^2 \cdot \frac{x}{2}\, dx = 2.$$

Applying the definitions of the previous slide, this gives,

$$\begin{aligned} \text{the mean} &= \mu = E(x) = 4/3 \\ \text{the variance} &= \sigma^2 = E(x^2) - [E(x)]^2 = 2 - 16/9 = 2/9 \\ \text{the standard deviation} &= \sigma = \sqrt{2/9} \approx 0.47 \end{aligned}$$

Loosely speaking, we say that the standard deviation is a measure of how close the graph of f is to the mean on average.

Median

Just as the mean measures the center of probability density function, the median m is the number such that half of the area under the curve is the left of m, and half the area under the curve is to the right of m. In symbols, this can be interpreted as the number m such that

$$\int_m^\infty f(x)\,dx = \frac{1}{2}.$$

Median

Just as the mean measures the center of probability density function, the median m is the number such that half of the area under the curve is the left of m, and half the area under the curve is to the right of m. In symbols, this can be interpreted as the number m such that

$$\int_m^\infty f(x)\,dx = \frac{1}{2}.$$

Solution: Applying this to the function $f(x) = \frac{x}{2}$ on $[0, 2]$, we look for the number m that satisfies

$$\int_m^2 \frac{x}{2}\,dx = \frac{1}{2}.$$

This gives

$$\int_m^2 \frac{x}{2}\,dx = \left.\frac{x^2}{4}\right|_m^2 = 1 - \frac{m^2}{4}.$$

For this quantity to equal 1/2 we must have $m = \sqrt{2}$. So we find the $\sqrt{2}$ is the median of this distribution. Remember the distribution is created by two things: a continuous random variable with an associated probability density function.

An Important PDF

Consider the function, $f(x) = e^{-x^2/2}$ over the real line, $(-\infty, \infty)$. The graph of this function is the famous bell curve, or *Gaussian curve*. The function has an anti derivative, but it is not in terms of any elementary function. (Recall from the first lecture this integral is given by erf(x), the error function.) It can be shown that the improper integral

$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}$$

so this function is not a probability density function, but

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is a probability density function called the ***standard normal distribution***.

Definition

A continuous random variable x is *normally distributed* with mean μ and standard deviation σ if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}.$$