Math 3070, Applied Statistics

Section 1

September 23, 2019

Lecture Outline, 9/23

Section 4.2

- Cumulative Distribution Functions
- Expected Value and Variance of a Continuous Random Variable

Preface

Most definitions for continuous random variables change $\sum$ to $\int$ and usually work the same way.

CDF of Continuous Random Variable

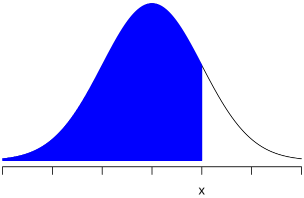
The **cumulative distribution function** (CDF), $F(x)$, of a continuous random variable X is defined the same as in the discrete case:

$$F(x) = P(X \leq x)$$

If X has pdf $f(x)$, then this becomes

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

By the Fundamental Theorem of Calculus, $F'(x) = f(x)$, if $F'(x)$ exists at x.



Example, CDF of $Unif(0,1)$

Compute the CDF of $X \sim unif(0,1)$.

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

When $x < 0$,

$$P(X < x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} 0dx = 0.$$

When $0 < x < 1$,

$$P(X < x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} 1dx = x.$$

When $1 < x$,

$$P(X < x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{1} 1dx = 1.$$

Example, CDF of $Unif(0, 1)$

Compute the CDF of $X \sim unif(0, 1)$.

$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x < 1 \\ 1, & 1 \le x \end{cases}$$

Properties of CDFs

Useful for calculation:

- $P(X > a) = 1 - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a)$

Useful for double-checking a function is a CDF:

- $\lim_{x \to -\infty} P(X < x) = 0$
- $\lim_{x \to \infty} P(X < x) = 1$
- CDFs of continuous random variables are continuous.

Percentiles and Median, Definition

Let p be a number between 0 and 1. The $(\mathbf{100p})^{\mathbf{th}}$ **percentile** of the distribution of a continuous random variable X is denoted by $\eta(p)$ and defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)dy$$

Alternatively, $\eta(p) = F^{-1}(p)$ if $F(x)$ is invertible. If it's not we usually take the smallest $\eta(p)$ that suffices. Won't need to consider that in this class.

The **median** $\tilde{u}$ of a continuous random variable X is the 50^{th} percentile or the percentile with $p = 0.5$.

This corresponds to the median of a data set. Roughly half of the observations will be below $\tilde{u}$.

Example, Median

Calculate the median of a random varible with the following PDF:

$$f(x) = \begin{cases} e^{-x+1} & \text{if } 1 \leq x \\ 0 & \text{otherwise} \end{cases}$$

When $x < 1$,

$$P(X < x) = \int_{-\infty}^{x} f(x)dx = 0$$

When $x \geq 1$,

$$P(X < x) = \int_{-\infty}^{x} f(x)dx = \int_{1}^{x} e^{-s+1}ds$$

$$= -e^{-s+1}\Big|_{s=1}^{x} = 1 - e^{-x+1}$$

Example, Median

Calculate the median of a random varible with the following PDF:

$$F(x) = \begin{cases} 1 - e^{-x+1}, & \text{if } 1 \leq x \\ 0, & x < 1 \end{cases}$$

$F(x) = 0.5$ when $x \geq 1$. Need to invert the function in that region.

$$\begin{aligned} 0.5 &= 1 - e^{-\tilde{u}+1} \\ 0.5 &= e^{-\tilde{u}+1} \\ \ln(0.5) &= -\tilde{u} + 1 \\ 1 - \ln(0.5) &= \tilde{u} \\ \tilde{u} &\approx 1.69315 \end{aligned}$$

Summary, Cumulative Density Function

- CDF: $F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt$
- $F'(x) = f(x)$ when $F'(x)$ exists.
- $P(X > a) = 1 - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a)$
- $\lim_{x \to -\infty} P(X < x) = 0$
- $\lim_{x \to \infty} P(X < x) = 1$
- CDFs of continuous random variables are continuous.
- $100p^{th}$ percentile $\eta(p)$: $p = F(\eta(p))$
- median $\tilde{u}$, $p = 0.5$ percentile

Expected Value, Definition

The **expected value** or **mean** of a continuous random variable X with PDF $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$

If $h(x)$ is a function then

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx.$$

Mean has the same interpretation as the discrete case or from data, a measure of center or location. And, it is also linear,

$$E[g(X) + ah(X) + b] = E[g(X)] + aE[h(x)] + b.$$

Why? Integrals are linear.

Variance, Definition

The **variance** of a continuous random variable X with PDF $f(x)$ and $E[X] = \mu$ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2].$$

The **standard deviation** (SD) of X is $\sigma_x = \sqrt{V(X)}$.

Same interpretation, average spread. Shorcut formula and linear transforms work the same too.

$$V(X) = E[X^2] - E[X]^2$$

$$V(aX+b) = a^2 V(X)$$

Variance, Derivations for Shortcut Formula and Linear Transforms

Shortcut Formula:

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx \\ &= \int_{-\infty}^{\infty} (x)^2 \cdot f(x)dx - 2\mu \int_{-\infty}^{\infty} x \cdot f(x)dx + \mu^2 \int_{-\infty}^{\infty} f(x)dx \\ &= E[X^2] - \mu^2 = E[X^2] - E[X]^2 \end{aligned}$$

Linear Transforms: Note: $E[aX+b] = aE[X] + b = a\mu + b$

$$\begin{aligned} V(aX+b) &= \int_{-\infty}^{\infty} [ax+b-(a\mu+b)]^2 \cdot f(x)dx \\ &= \int_{-\infty}^{\infty} a^2[x-\mu]^2 \cdot f(x)dx \\ &= a^2 V(X) \end{aligned}$$

Example, Mean and Variance of the a Uniform RV

Compute the mean and variance of $X \sim unif(a, b)$.

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x < b \\ 0, & \text{otherwise} \end{cases}$$

guess.

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx = \int_a^b \frac{x}{b-a}dx \\ &= \frac{x^2}{2(b-a)}\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2} \end{aligned}$$

Example, Mean and Variance of the a Uniform RV

Compute the mean and variance of $X \sim unif(a, b)$.

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x < b \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned}
V(X) &= E[X^2] - E[X]^2 \\
&= \int_a^b \frac{x^2}{b-a} dx - \left(\frac{b+a}{2}\right)^2 = \frac{1}{b-a} \frac{x^3}{3}\bigg|_{x=a}^{b} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{(a^2+ab+b^2)}{3} - \frac{b^2+2ba+a^2}{4} \\
&= \frac{(4a^2+4ab+4b^2)}{12} - \frac{3b^2+6ba+3a^2}{12} \\
&= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}
\end{aligned}$$

Example, Mean and Variance of the a Uniform RV

Compute the mean and variance of $X \sim unif(a, b)$.

$$E[X] = \frac{b+a}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$

Takeaway:

- The mean is the average of the end points.
- The variance is explictly related to the distance between the endpoints $b - a$.

Example, Modeling with a Uniform RV

A random number generator produces values that follow uniform random variable. Researchers take find a sample mean of 5 and a sample standard deviation of $\sqrt{12}$. Determine the minimum and maximum values assuming that the sample mean and variance are the true mean and standard deviation.

Using what was found in the previous problem,

$$\frac{b+a}{2} = 5 \text{ and } \frac{b-a}{\sqrt{12}} = \sqrt{12}$$

or

$$b + a = 10 \text{ and } b - a = 12$$

Using linear algebra,

$$b = 11 \text{ and } a = -1.$$

Maximum value $= 11$ and minimum value $= -1$.

Example, Modeling with a Uniform RV

A random number generator produces values that follow uniform random variable. Researchers take find a sample mean of 5 and a sample standard deviation of $\sqrt{12}$. Determine the minimum and maximum values assuming that the sample mean and variance are the true mean and standard deviation.

Closing note, $P(X = b) = P(X = a) = 0$ or it is impossible to observe the endpoints of a uniform random variable. Maximum and minimum values of the data set may not work as well as the sample mean and variance.

Summary, Expected Value and Variance

- $\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$
- $E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$
- $E[g(X) + ah(X) + b] = E[g(X)] + aE[h(x)] + b$
- $\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx = E[(X - \mu)^2]$
- $V(X) = E[X^2] - E[X]^2$
- $V(aX + b) = a^2 V(X)$
- If $X \sim unif(a, b)$

$$E[X] = \frac{b + a}{2} \quad V(X) = \frac{(b - a)^2}{12}$$