Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 3 (continued)

- Generalize distributions to the continuous case
- Density functions
- Expected value, variance, standard deviation on these models
- How to compute these parameters

Beyond Discrete Models

A **discrete random variable** takes on values with spaces in between them.

- Geometric: $Geom(p)$. $X \in \{1, 2, 3, \ldots\}$
- Binomial: $Binom(n, p)$. $X \in \{0, 1, 2, \ldots, n\}$
- Poisson: $Poisson(\lambda)$. $X \in \{0, 1, 2, \ldots\}$.
- Negative Binomial: $NegBinom(k, p)$ $X \in \{k, k+1, k+2, \ldots\}$

A **continuous random variable** is a random quantity that can take on any value on a continuous scale (a smooth interval of possibilities). Examples:

- Amount of water you drink in a day
- How long you wait for a bus
- How far you live from the nearest grocery store

Some Awkward Questions

How do I make a probability table for such a situation?

What is the probability of drinking exactly 1.759823 liters in a given day?

Is the denominator of

$$P(A) = \frac{\#\text{ outcomes in event}}{\#\text{ outcomes in sample space}}$$

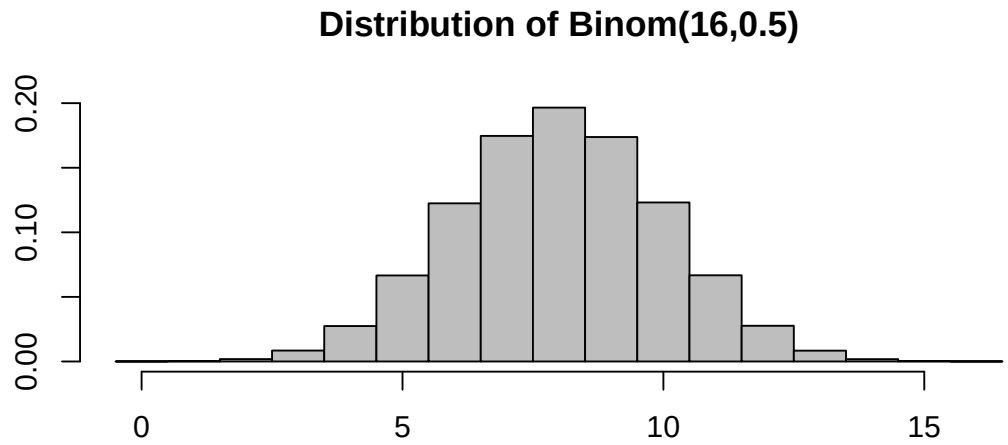
equal to ∞?

For continuous random variables, we have to:

- change how we present the probability model (no more tables!)
- alter how we ask questions
- generalize our definition of probability
- rethink what a probability of 0 means

From Discrete to Continuous Random Variables

Suppose you flip a fair coin $n = 16$ times and record how many Heads you get. $X = Binom(16, 0.5)$.

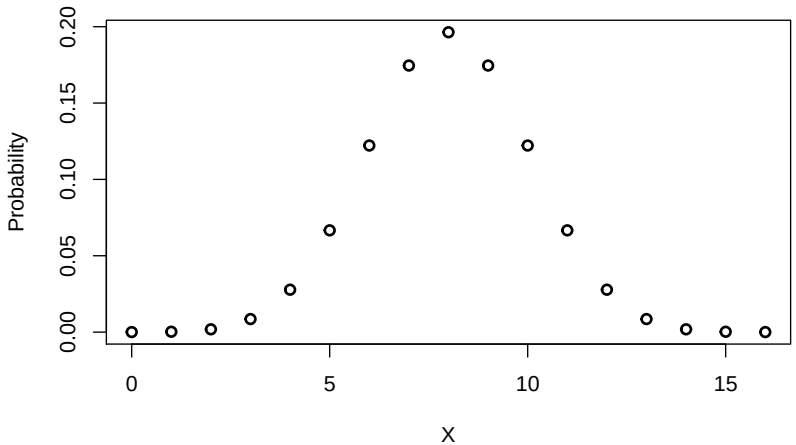


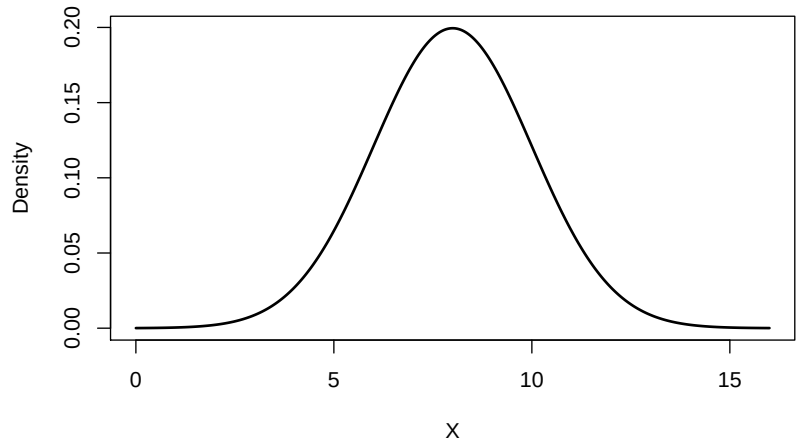
When we visualize the probability table:

- An outcome is more likely if there is more area in the rectangle for that value
- The sum of the areas of the bars must be 1
- The bar heights must be at least 0 (no negative heights)

Taking the leaps!

We momentarily use dots insteads of rectangles. The picture suggests how we could generalize: draw curves instead of dots.





With discrete models, two ideas are linked to probability:

- Height of the y-axis
- Areas of rectangles

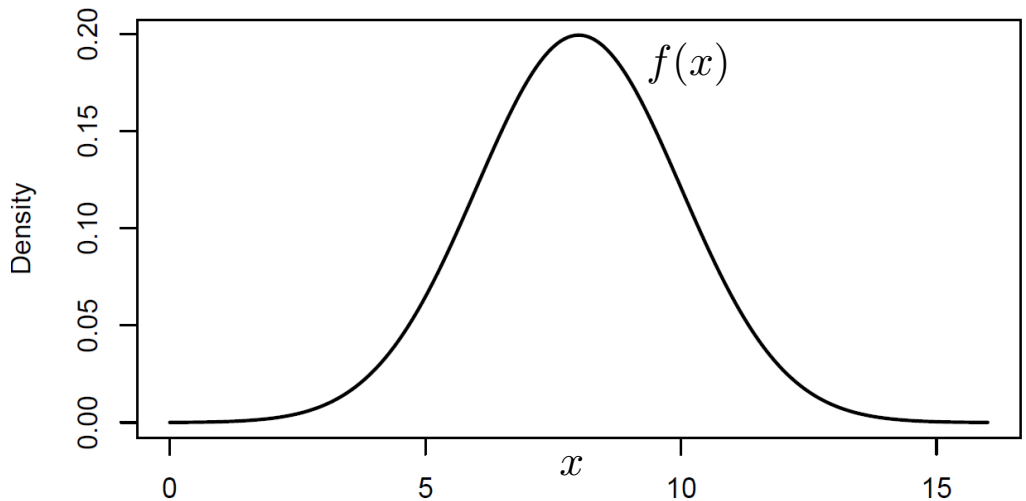
With continuous random variables probability is linked to:

- Areas only

Building the New Universe

We model a continuous random variable X through a **density function** $f(x)$ which has two properties:

- $f(x) \geq 0$ for all x (heights must be at least 0)
- $\int_{\infty}^{\infty} f(x)dx = 1$ (sum of areas must be 1)



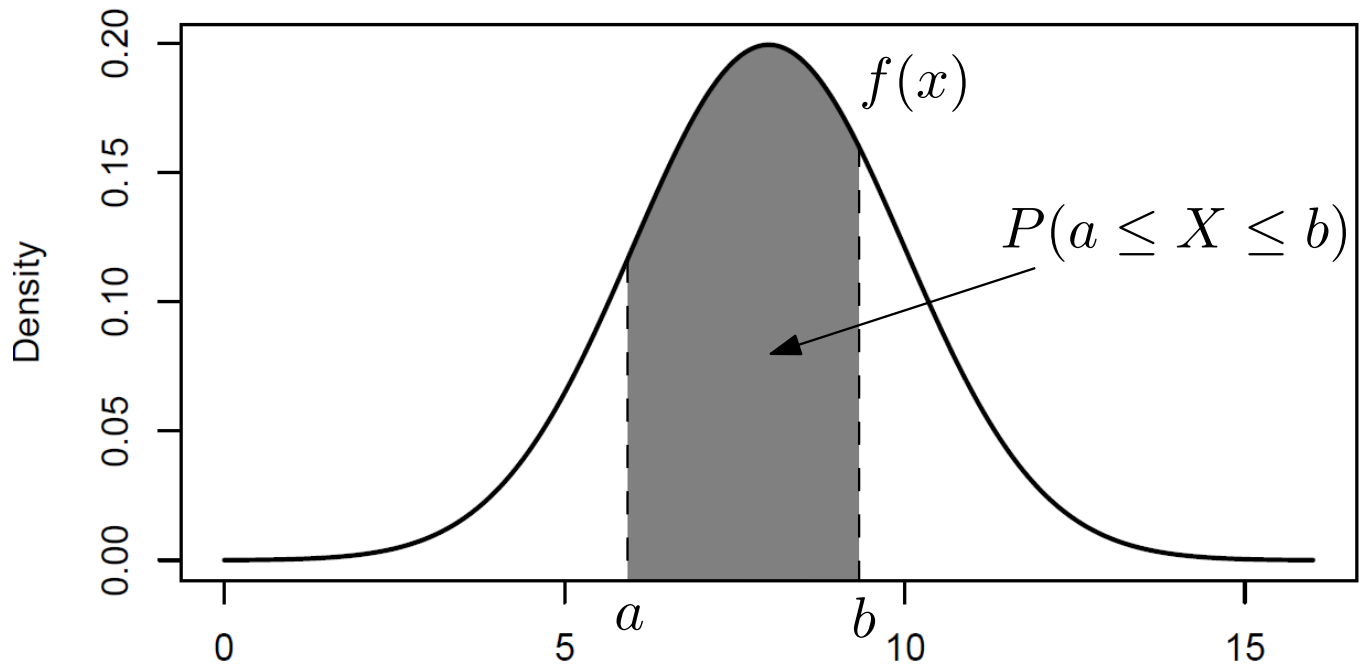
A higher value of $f(x)$ means values nearby x are more likely.

Key Definition

If $f(x)$ is a density function for the continuous random variable X, then we define

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

(This formalizes that probability is related to areas)



Examples

Once we are told what $f(x)$ is:

To find the probability a person drinks between 2 and 3 liters of water a day, we evaluate

$$\int_2^3 f(x)dx.$$

To find the probability a person drinks more than 1 liter of water a day, we evaluate

$$\int_1^\infty f(x)dx.$$

To find the probability a person drinks less than 4 liters of water a day, we evaluate

$$\int_{-\infty}^4 f(x)dx.$$

The Counter-Intuitive Continuous World

The probability of any particular outcome happening is 0. Indeed,

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)dx = 0.$$

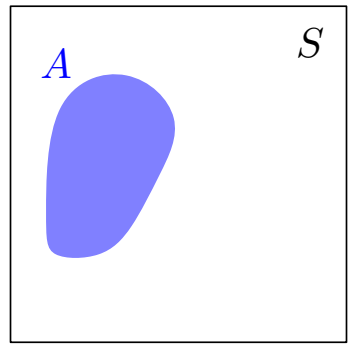
The density graph $f(x)$ is not $P(X = x)$! It is a function that helps you figure out probabilities by examining the area underneath it. Its shape suggest that values are more likely (relatively), buth the probability of any particular outcome is still 0.

You ask questions about probability of some interval of values occurring since the probability of any individual outcome is always 0.

Probabilities = Areas

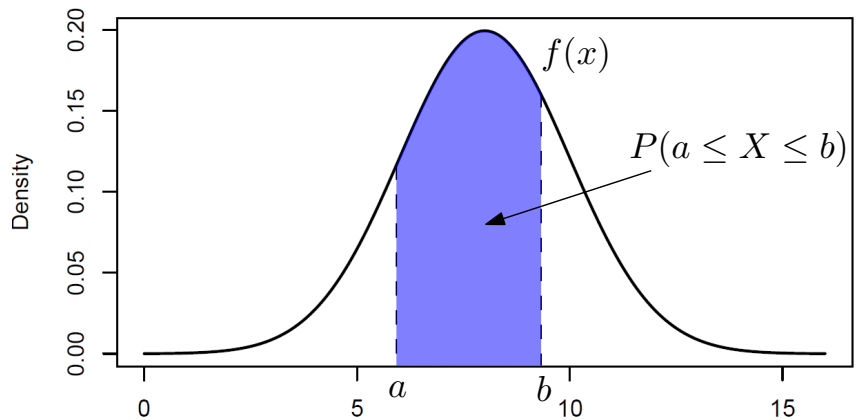
You can reinterpret the discrete world with areas too:

$$P(A) = \frac{\#\text{ outcomes in A}}{\#\text{ outcomes in S}} = \frac{Area(A)}{Area(S)}.$$



Which should sound similar to

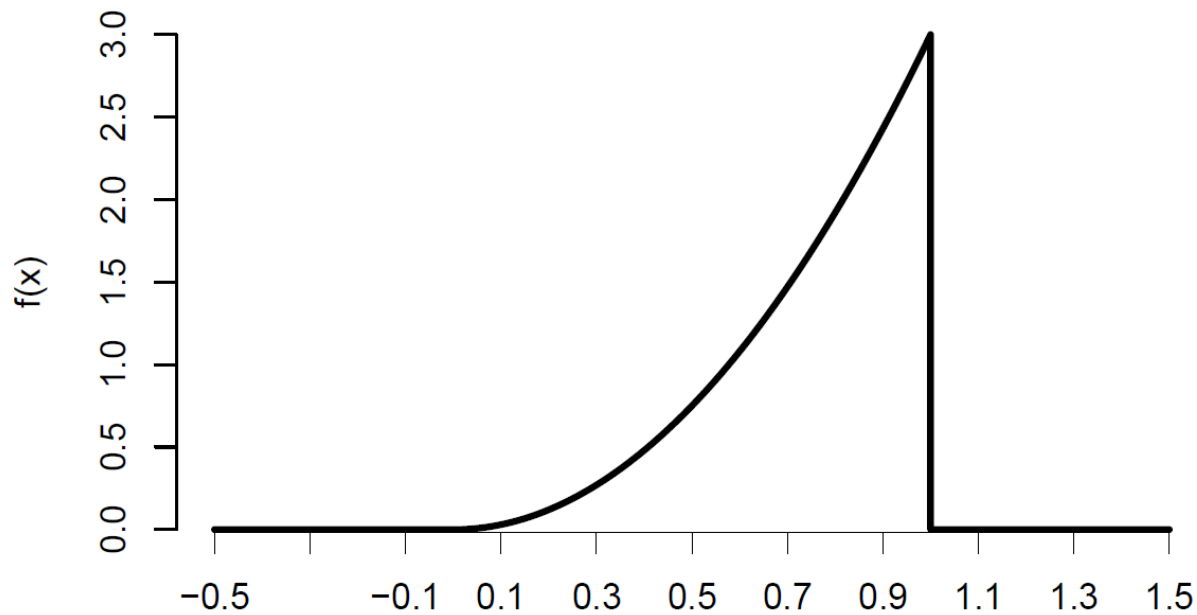
$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$



Example

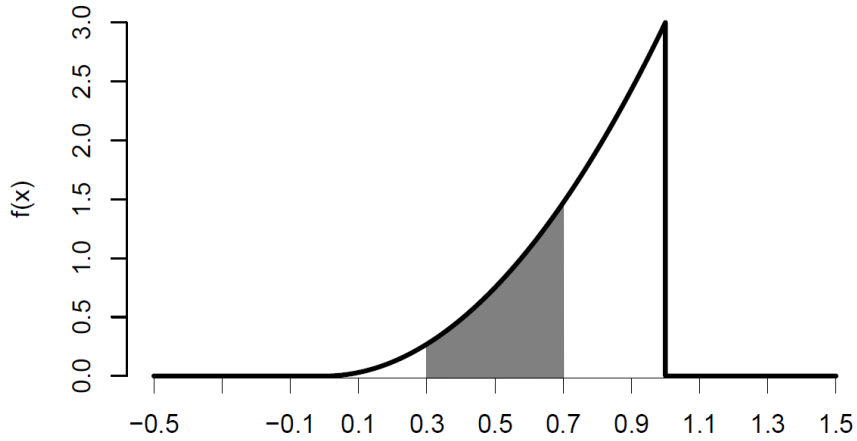
Suppose the concentration of iodine in a chemical sample is modeled by the density function

$$f(x) = \begin{cases} 3x^2 & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

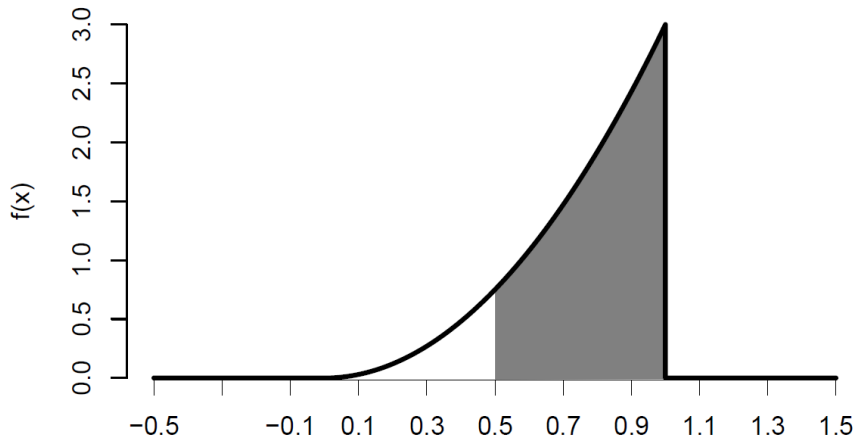


Example

$$
\begin{aligned}
P(0.3 \leq X \leq 0.7) &= \int_{0.3}^{0.7} f(x)dx \\
&= \int_{0.3}^{0.7} 3x^2 dx \\
&= x^3\big|_{0.3}^{0.7} \\
&= (0.7)^3 - (0.3)^3 \\
&= 0.316.
\end{aligned}
$$



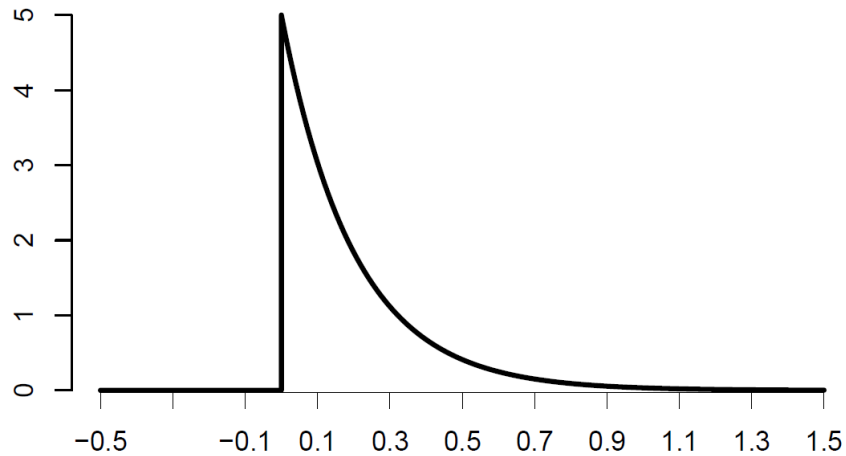
$$
\begin{aligned}
P(X \geq 0.5) &= \int_{0.5}^{\infty} f(x)dx \\
&= \int_{0.5}^{1} 3x^2 dx \\
&= x^3\big|_{0.5}^{1} \\
&= 1^3 - (0.5)^3 \\
&= 0.875.
\end{aligned}
$$



Example

Suppose the distance a freshman lives from UCSD (in miles) is modeled by the density

$$g(x) = \begin{cases} 5e^{-5x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$



Show that $g(x)$ is a valid density function.

We have $g(x) \geq 0$ for all x, and

$$\begin{aligned} \int_{-\infty}^{\infty} g(x)dx = \int_0^{\infty} 5e^{-5x}dx = \lim_{n\to\infty} \int_0^n 5e^{-5x}dx &= \lim_{n\to\infty} -e^{-5x}\Big|_0^n \\ &= \lim_{n\to\infty} (-e^{5n} - (-e^0)) \\ &= 1. \end{aligned}$$

Example

What is the probability that a freshman lives exactly 3 miles from school?

This is the wrong kind of question to ask with a continuous random variable. We immediately know the answer: $P(X = 3) = 0$.

What is the probability a freshman lives less than 1 mile from school?

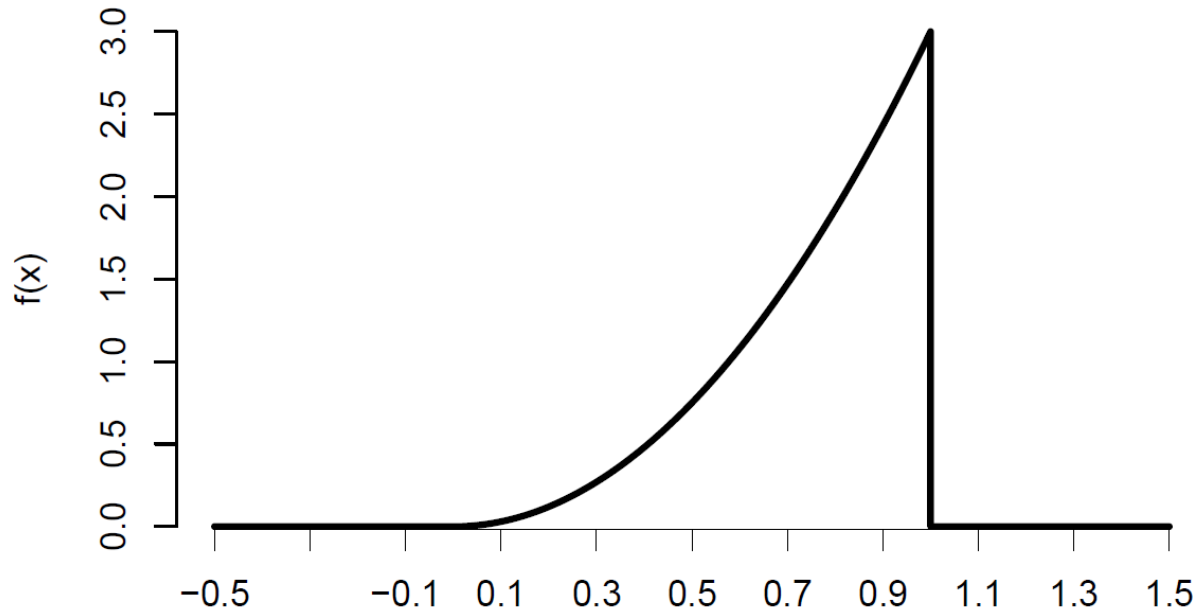
$$P(0 \leq X < 1) = \int_0^1 5e^{-5x}dx = -e^{-5x}\Big|_0^1 = -e^{-5} + e^0 = 1 - e^{-5} \simeq 0.993.$$

Remark: Here, inclusive or strict inequalities ($\leq$ or $<$) don't make any difference. Indeed, for continuous random variables,

$$\begin{aligned} P(a \leq X \leq b) &= P(a \leq X < b) + P(X = b) \\ &= P(a \leq X < b) + 0 \\ &= P(a \leq X < b) \end{aligned}$$

Example

How far, on average, do we expect a freshman would live from campus?



This is not asking for a probability, but for a fact related to the entire model.
We need to know the expected value of a continuous random variable!

Expected Value in the Continuous Case

For a discrete random variable, we had $E(X) = \sum_x xP(X = x)$.

For a continuous random variable, the density function $f(x)$ helps us find probabilities, so we define

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Remark: Going from discrete to continuous:

- The sum becomes an integral
- We replace the probability function with the density function

In the previous example, the expected distance is

$$E(X) = \int_{-\infty}^{\infty} x\left(5e^{-5x}\right)dx = \ldots$$

Variance in the Continuous Case

For a discrete random variable, $Var(X) = \sum_x (x-\mu)^2 P(X=x)$.

For a continuous random variable, the density function $f(x)$ helps us find probabilities, so we define

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

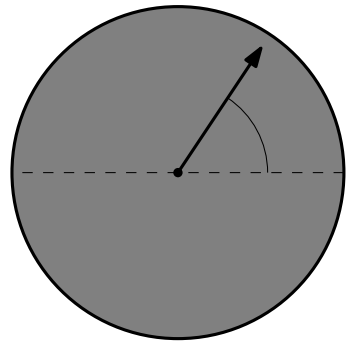
Easier version:

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= E(X^2) - \mu^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2. \end{aligned}$$

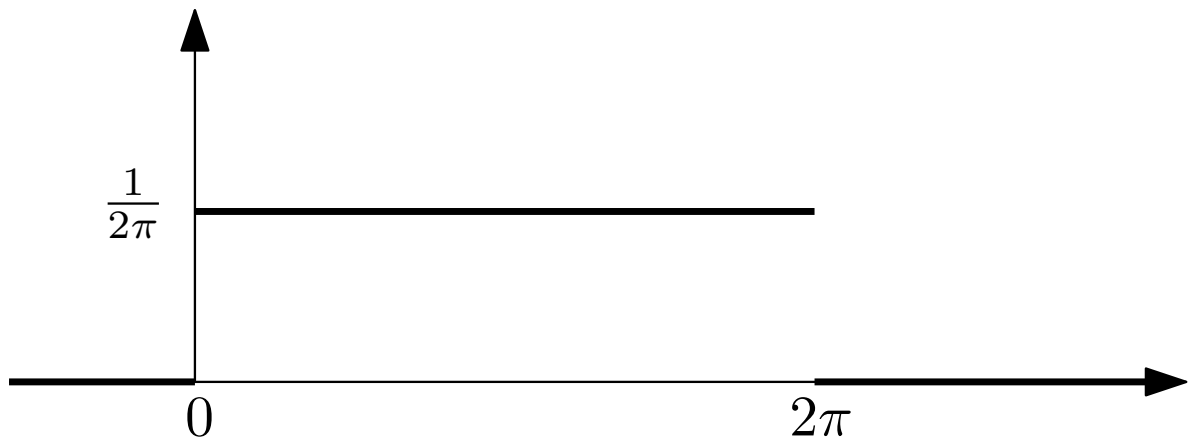
The standard deviation is $SD(X) = \sqrt{Var(X)}$.

Flicking a Spinner

You flick a spinner and measure the angle it makes (from horizontal), in radians, between 0 and 2π. Assuming each angle is equally likely, find the average angle and standard deviation.

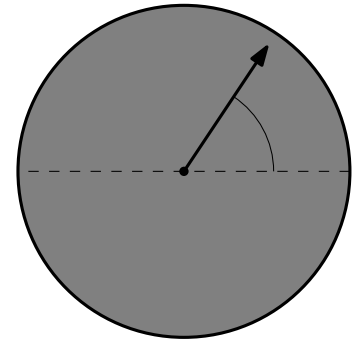


We must create our own density function!



The correct height is $\frac{1}{2\pi}$, since it makes a rectangle of area 1.

Flicking a Spinner: Parameters



$$\text{Let } f(x) = \begin{cases} \frac{1}{2\pi} & 0 \le x \le 2\pi \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{2\pi} \int_0^{2\pi} x dx$$

$$= \frac{1}{2\pi} \left. \frac{x^2}{2} \right|_0^{2\pi} = \frac{1}{2\pi} \frac{4\pi^2}{2} = \pi \text{ rad.}$$

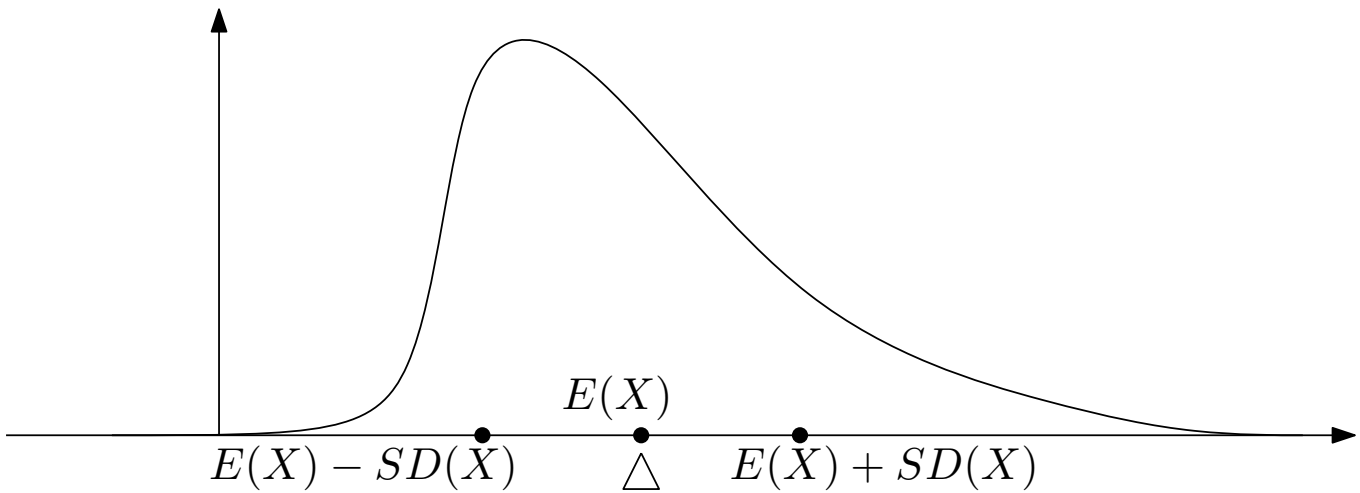
$Var(X) = E(X^2) - E(X)^2$, and

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{2\pi} \int_0^{2\pi} x^2 dx$$

$$= \frac{1}{2\pi} \left. \frac{x^3}{3} \right|_0^{2\pi} = \frac{1}{2\pi} \frac{4\pi^2}{3}$$

$$Var(X) = \frac{4\pi^2}{3} - \pi^2 = \frac{\pi^2}{3} \text{ , and } SD(X) = \sqrt{Var(X)} = \frac{\pi}{\sqrt{3}} \text{ rad.}$$

Visual Interpretation of the Mean and Standard Deviation



The mean $E(X)$ is the balance point of the density function.

The standard deviation $SD(X) = \sqrt{Var(X)}$ is the distance you must go from the mean to embrace a large chunk of the most likely outcomes. It gives a sense of how compressed the possibilities are around the mean.

Practice

After serious investigations, a TA claims that the free time she has before a student arrives (in minutes) in a given office hour is given by the density:

$$f(x) = \begin{cases} \dfrac{60 - x}{1800} & 0 \leq x \leq 60 \\ 0 & \text{otherwise.} \end{cases}$$

Is it a valid density function ?

Find the mean arrival time of that first student, and the standard deviation.