Unit #3: Probability Distributions and Probability Density Functions

(Ch 3.4, 3.4, 4.1, 4.2, 4.3)

Learning Objectives

At the end of this unit, students should be able to:

1. Distinguish between a *continuous* and *discrete* random variable.
2. Distinguish between a random variable and a *realization* of a random variable.
3. Define a probability mass function (discrete) and a probability density function (continuous) for a rv X.
4. Articulate and justify properties of pmfs and pdfs.
5. Calculate probabilities using pdfs/pmfs.
6. Identify situations for which a Bernoulli, binomial, geometric, or Poisson distribution works as a good model.
7. Calculate the probability that a Bernoulli, Binomial, Negative Binomial, Geometric, or Poisson rv takes on a particular value or set of values.
8. Identify situations for which a uniform, exponential, or beta distribution works as a good model.
9. Calculate the probability that a uniform, exponential, or beta rv falls within a given range.
10. Define the *cumulative distribution function (cdf)* for a rv. Calculate the cdf for given values of x.

Random Variables

Reminder: A *random variable* assigns a number to every event associated with a probabilistic process.

The rv takes on a finite or countably infinite set of values

Random variables can be **discrete** or **continuous.**

r.v. takes on values in a continuum.

Examples:

X = # of heads in 3 flips.

Y = points scored during a football game

Z = time waiting at a bus stop

Big Picture: In statistics, we will model *populations* using random variables, and features/parameters (e.g., mean, variance) of these random variables will tell us about populations.

Random Variables

Important Notation: Random variables are usually denoted by uppercase letters near the end of our alphabet (e.g. *X, Y*).

The value that a random variable takes on is usually denoted by a lowercase letter, such as *x,* which correspond to the r.v. *X.*

Examples:

$$\underbrace{P(\underset{\text{r.v.}}{X} = \underset{\#}{x})}, \quad \underbrace{P(Z \geq z), \quad P(a \leq Y \leq b)}_{\text{rv}}$$

Probability Density/ Distribution Functions

Definition: A *Probability distribution function (pdf)* is a function that describes the probability distribution of a random variable X.

If X is **discrete**, the pdf provides answers to questions like $\underline{P(X=x)}$. It is also called a *probability mass function (pmf).*

If X is **continuous**, then $\underline{P(X=x)}$ = 0 for ***all*** x. Why?! In this case, the distribution function is called a *probability density function*. (pdf)

In the continuous case, the pdf provides answers to questions like:

$P(X \le x)$, $P(a \le X \le b)$

Properties of pdfs

For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x.

2. (a) For discrete distributions: $\sum_x P(X=x) = 1$

   (b) For continuous distributions: $\int_x f(x)\,dx = 1$

Discrete Random Variables

Probability Distributions (Discrete)

The probability distribution function (pdf) _______ of a discrete r.v. *X* describes how the total probability is distributed among all the possible range values of the r.v. *X*. That is:

Example

A lab has 6 computers. Let *X* denote the number of these computers that are in use during lunch hour -- {0, 1, 2… 6}.

Suppose that the probability distribution of *X* is as given in the following table:

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Example

From here, we can find many things:

1. Probability that at most 2 computers are in use

2. Probability that at least half of the computers are in use

3. Probability that there are 3 or 4 computers free

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Discrete random variables can be categorized into different types or classes. Each type/class models many different real-world situations.

The Bernoulli Distribution/RV

Bernoulli random variable: Any random variable whose only possible values are 0 or 1.

This is a discrete random variable – why?

This distribution is specified by a single parameter:

The Bernoulli Distribution/RV

Bernoulli random variable: Any random variable whose only possible values are 0 or 1.

Examples of real-world scenarios modeled by a Bernoulli rv:

<u>NOTATION</u>: We write $X \sim Ber(p)$ to indicate that X is a Bernoulli rv with success probability p.

The Binomial Distribution/RV

Motivating Example: A fair coin is tossed four times. A "successful toss" is defined to be the coin landing on heads. Let X = # of successes/heads in 4 tosses.

What is *P(X = 2)*?

What is *P(X = 3)*?

The Binomial Distribution/RV

Let's use the probabilities we calculated in the example above to derive the binomial pdf.

<u>NOTATION</u>: We write $\underline{X \sim Bin(n,p)}$ to indicate that *X* is a binomial rv based on *n* Bernoulli trials with success probability *p.*

The Binomial Distribution/RV

The Binomial r.v. counts the **total number of successes** out of n trials, where X is the number of successes.

Important Assumptions:

Each trial must be independent of the previous experiment.

The probability of success must be the same for each trial.

The Geometric Distribution/ RV

Motivating example: A patient is waiting for a suitable matching kidney donor for a transplant. The probability that a randomly selected donor is a suitable match is 0.1.

P('1st donor matches') = 0.1; P('2nd donor matches') = (0.9)(0.1)

What is the probability the first donor tested is the first matching donor? Second? Third?

P('3rd donor matches) = $(0.9)^2(0.1)$

The Geometric Distribution/ RV

\# of failures until the 1st success.

Continuing in this way, a general formula for the pmf emerges:

$$P(X = x) = (1-p)^x p$$

\# of failures

The parameter *p* can assume any value between 0 and 1. Depending on what parameter *p* is, we get different members of the ***geometric*** distribution.

NOTATION: We write $X \sim G(p)$ to indicate that *X* is a geometric rv with success probability *p.*

The Negative Binomial Distribution/RV

Motivating Example: A fair coin is tossed four times. A "successful toss" is defined to be the coin landing on heads. Let X = # of failures/tails before the second success/heads.

binomial

r^{th} success

F F F F S S F | S

$P(X=x)$, $X \sim Bin(n-1, p)$

How is this related to the geometric distribution? The binomial distribution?

It's a generalization
(stop at r^{th} success
where geo, $r=1$)

${}_nC_k = \binom{n}{k}$

$Y \sim Bin(n,p)$, $P(Y=y) = \binom{n}{y} p^y (1-p)^{n-y}$, $y \in \{0,\dots,n\}$

The Negative Binomial Distribution/RV

In general, let X = $\underbrace{\text{\# of failures before the } r^{th} \text{ success}}_{x}$. The pdf/pmf is:

$$P(X=x) = \underbrace{\binom{x+r-1}{r-1} p^{r-1} (1-p)^{[x+r-1]-[r-1]}}_{Bin(x+r-1,\,p)}\, p$$

$$= \binom{x+r-1}{r-1} p^r (1-p)^x, \quad x \in \{0,1,\dots\}$$

NOTATION: We write $X \sim NB(r,p)$ to indicate that *X* is a negative binomial r.v., with *x* failures occurring before *r* successes, where the probability of success is equal to p.

$\boxed{\text{In R: \_nbinom( )}}$

Example

A physician wishes to recruit 5 people to participate in a new health regimen. Let p = .2 be the probability that a randomly selected person agrees to participate. What is the probability that 15 people must be asked before 5 are found who agree to participate? X = # of people who $\underbrace{\text{refuse}}_{\text{failure}}$ before $\underset{\uparrow \, r}{5}$ people $\underbrace{\text{agree}}_{\text{success}}$

$$P(X=10) = \binom{10+5-1}{5-1}(0.2)^5(0.8)^{10} = 0.0344$$

The Poisson Distribution/RV

A Poisson r.v. describes the total number of events that happen in a certain <u>time period</u>.

Examples:

# of vehicles arriving at a parking lot in one week

# of gamma rays hitting a satellite per hour

# of cookies sold at a bake sale in 1 hour

The Poisson Distribution/RV

A Poisson r.v. describes the total number of events that happen in a certain time period.

A discrete random variable X is said to have a **Poisson distribution** with parameter λ ($\lambda > 0$) if the pdf of X is

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

NOTATION: We write $X \sim \mathcal{P}(\lambda)$ to indicate that X is a Poisson r.v. with parameter λ.

$\text{Pois}(\lambda)$

The Poisson Distribution/RV

$X \sim \mathcal{P}(\lambda)$. $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, 2, \ldots$

Example: Let *X* denote the number of mosquitoes captured in a trap during a given time period. Suppose that *X* has a Poisson distribution with λ = 4.5. What is the probability that the trap contains 5 mosquitoes?

$$P(X = 5) = \frac{(4.5)^5 e^{-4.5}}{5!} \approx 0.171$$

Example

A factory makes parts for a medical device company. 6% of those parts are defective. For each one of the problems below:
(i) Define an appropriate random variable for the experiment.
(ii) Give the values that the random variable can take on.
(iii) Find the probability that the random variable equals 2.
(iv) State any assumptions you need to make.

1. Out of 10 parts, x are defective.

2. Upon observing an assembly line, x non-defective parts are observed before finding a defective part.

3. The number of defective parts made per day, where the rate of defective parts per day is 10.

Example

A factory makes parts for a medical device company. 6% of those parts are defective. For each one of the problems below:
(i) Define an appropriate random variable for the experiment.
(ii) Give the values that the random variable can take on.
(iii) Find the probability that the random variable equals 2.
(iv) State any assumptions you need to make.

1. Out of $\overset{=n}{10}$ parts, x are defective.

(i) $X \sim Bin(10, 0.06)$

(ii) $X \in \{0, 1, 2, \dots, 10\}$

(iii) $P(X=2) = \underbrace{\binom{10}{2}}_{= {}_{10}C_2}(0.06)^2(0.94)^8 = 0.0988\dots$

(iv) · Trials are independent
· $p = 0.06$ is the same for all trials.

$P(X=x)=(1-p)^x p$

\# of failures

Example

A factory makes parts for a medical device company. 6% of those parts are defective. For each one of the problems below:
(i) Define an appropriate random variable for the experiment.
(ii) Give the values that the random variable can take on.
(iii) Find the probability that the random variable equals 2.
(iv) State any assumptions you need to make.

2. Upon observing an assembly line, x non-defective parts are observed before finding a defective part.

(i) $X \sim G(0.06)$

(ii) $X \in \{0,1,2,\ldots\}$

(iii) $P(X=2)=(0.94)^2(0.06) \approx 0.053\ldots$

(iv) Same as 1.

Example

A factory makes parts for a medical device company. 6% of those parts are defective. For each one of the problems below:
(i) Define an appropriate random variable for the experiment.
(ii) Give the values that the random variable can take on.
(iii) Find the probability that the random variable equals 2.
(iv) State any assumptions you need to make.

3. The number of defective parts made per day, where the rate of defective parts per day is 10.

(iii) $P(X=2) \approx 0.002$

(iv) Given an "exposure period"

r.v.s

D C

?

classes

pmf cdf

Cumulative Distribution Functions (cdf)

Definition: The cumulative distribution function (*cdf*) is denoted with $F(x)$.

For a discrete *r.v.* X with pdf $f(x) = P(X = x)$, the *cdf*, $F(x)$, is defined for every real number x to be the probability that the observed value of X will be at most x. Mathematically:

$$P(X \le x) = \sum_{k=0}^{x} P(X = k)$$

Cumulative Distribution Functions (cdf)

Suppose we are given the following pmf:

$$P(X=x) = p(x) = \begin{cases} .500 & x = 0 \\ .167 & x = 1 \\ .333 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

1. Calculate: F(0), F(1), F(2).

$F(0) = P(X \le 0) = P(X=0) = 0.5$; $F(1) = P(X \le 1) = P(X=0) + P(X=1) = 0.667$

$F(2) = 1$

3. What is F(1.5)? F(20.5)?

$F(1.5) = P(X \le 1.5) = F(1) = 0.667$

$F(20.5) = 1$

5. Is P(X < 1) = P(X <= 1)? No!

Show that the CDF is a non-decreasing function.

When is $F(x_1) = F(x_2)$?

Continuous Random Variables

Continuous Distributions/ RVs

Definition: A random variable *X* is *continuous* if possible values comprise either a single interval on the number line or a union of disjoint intervals.

Note: If *X* is continuous, P(X = x) = 0 for any x!

Example: If in the study of the ecology of a lake, *X*, the r.v. may be depth measurements at randomly chosen locations. $X \in [a, b]$, $a, b \in \mathbb{R}$

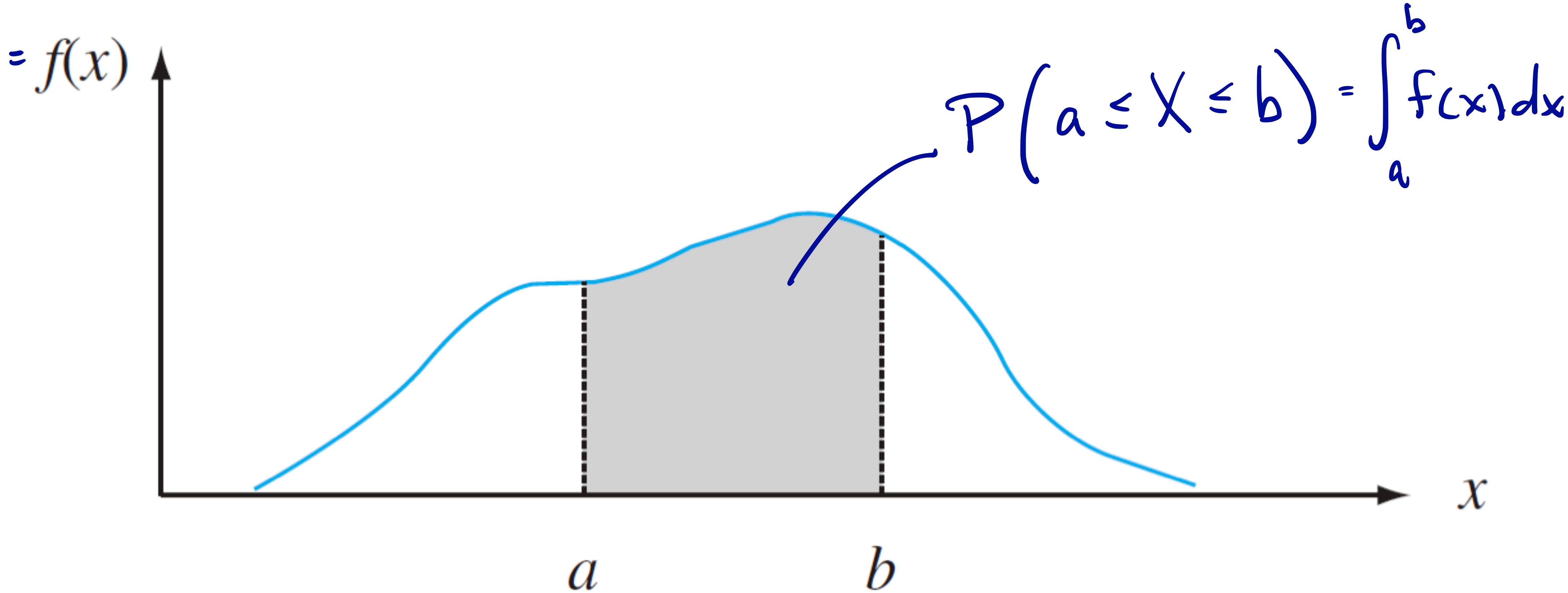
Other examples? Duration of lunch = Y

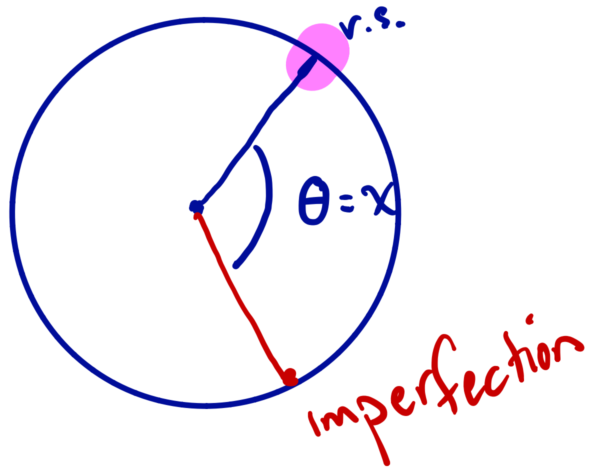
Z = waiting time for the bus.

Continuous Distributions/ RVs

The probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function:

pdf





Example

Example: Consider the reference line connecting the valve stem on a tire to the center point.

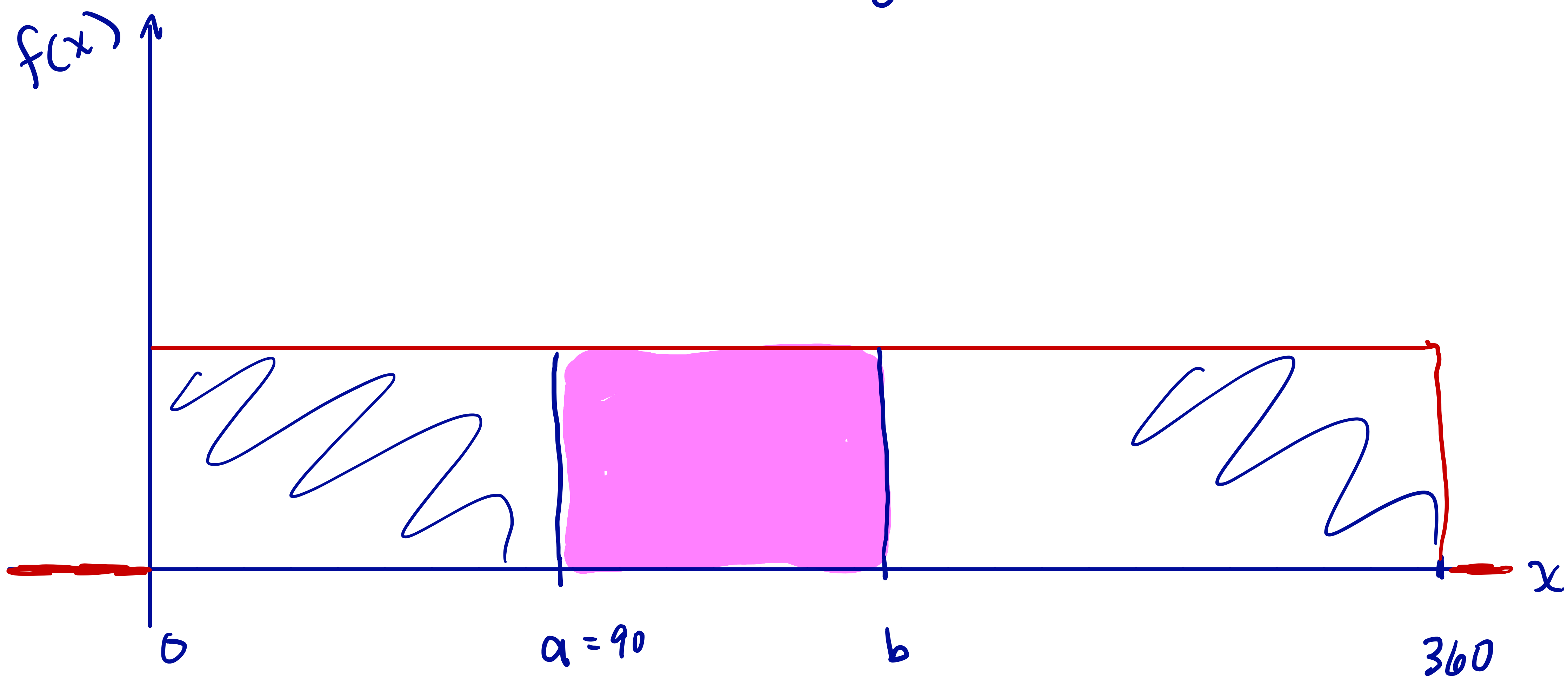
Let X be the angle measured clockwise to the location of an imperfection. The pdf for X is

$$f(x) = \begin{cases} \dfrac{1}{360} & 0 \leq x < 360 \\ 0 & \text{otherwise} \end{cases}$$

Example (Continued)

Graphically, the pdf is:

$$\int_0^{360} \frac{1}{360}\,dx = \frac{x}{360}\Big|_0^{360} = 1 - 0 = 1$$



$$P(a \le X \le b) = \int_a^b \frac{1}{360}\,dx$$

Example (Continued)

Clearly $f(x) \geq 0$. How can we show that the area of this pdf is equal to 1?

See last slide

How do we calculate P(90 <= X <= 180)?

$$\int_{90}^{180} \frac{1}{360}\,dx = \frac{1}{4}$$

What is the probability that the angle of occurrence is within 90 of the reference line? (The reference line is at 0 degrees.)

$$P(0 \leq X \leq 90) + P(270 \leq X < 360) = \frac{1}{2}$$

The Uniform Distribution/RV

The previous problem was an example of the **uniform distribution**.

Definition: A continuous rv X is said to have a *uniform distribution* on the interval [a, b] if the pdf of X is:

$$\underbrace{f(x;}_{\text{variable}}\underbrace{a,b}_{\text{parameters}}) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{o/w} \end{cases}$$

NOTATION: We write $X \sim U(a, b)$ to indicate that X is a uniform rv with a lower bound equal to a and an upper bound equal to b.

The Exponential Distribution/ RV

The family of exponential distributions provides probability models that are very widely used in engineering and science disciplines to describe **time-to-event data**.

Examples:

① X = time waiting for bus

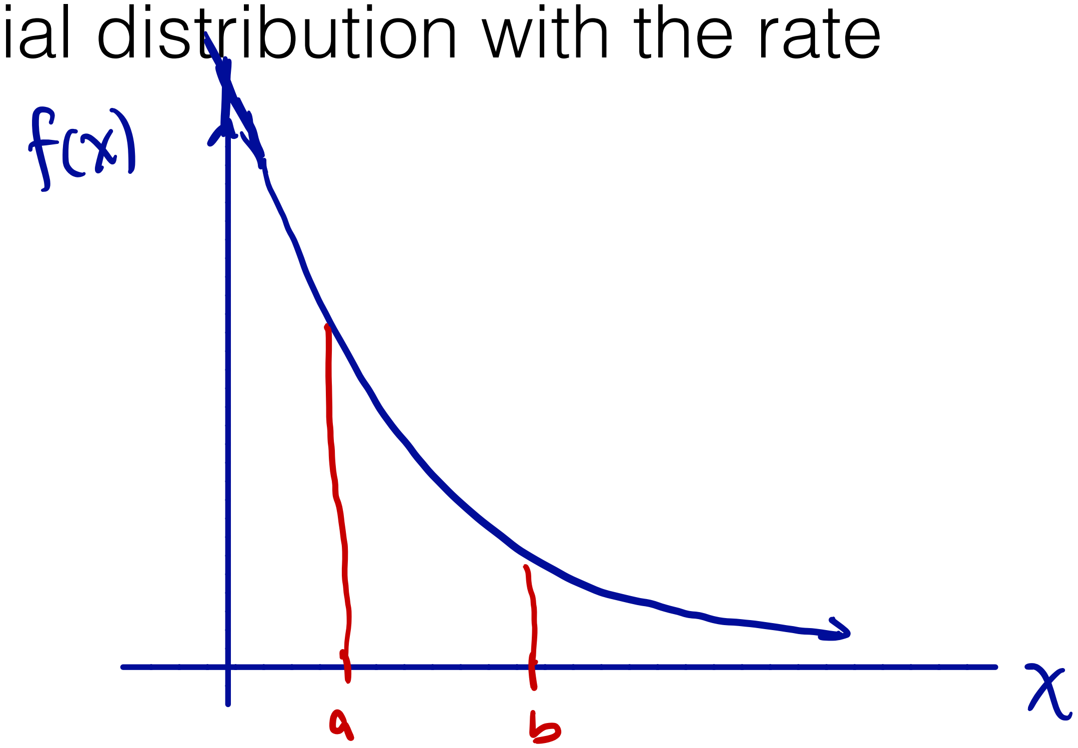
② Y = time until defective part is produced

The Exponential Distribution/ RV

The family of exponential distributions provides probability models that are very widely used in engineering and science disciplines to describe **time-to-event data**.

Definition: X is said to have an exponential distribution with the rate parameter ($\lambda > 0$) if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o/w} \end{cases}$$



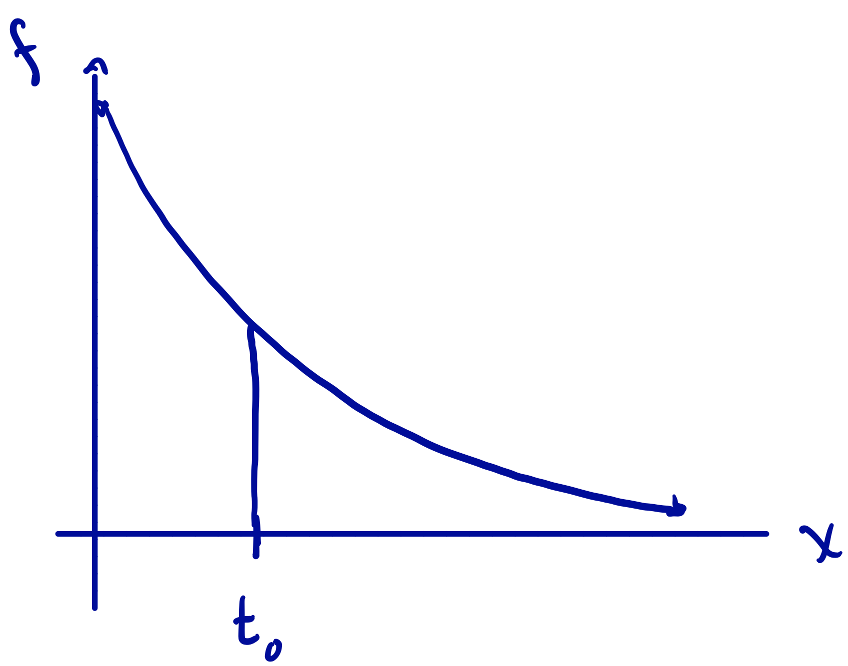
NOTATION: We write $X \sim Exp(\lambda)$ to indicate that X is an Exponential r.v. with parameter λ.

$\lambda > 0$

The Exponential Distribution/RV

$$\boxed{P(X > t_0) = \int_{t_0}^{\infty} \lambda e^{-\lambda x}\,dx =} e^{-\lambda t_0}$$

A partial reason for the popularity of such applications is the **memoryless property** of the Exponential distribution:



Let $X \sim Exp(\lambda)$. Suppose we know that $X > t_0$. what is the prob. that $X > t + t_0$?

$$P(X > t + t_0 \mid X > t_0) = P(X > t)$$

Proof:

$$P(X > t + t_0 \mid X > t_0) = \frac{P(X > t + t_0 \cap X > t_0)}{P(X > t_0)}$$

$$= \frac{P(X > t + t_0)}{P(X > t_0)} = \frac{\exp\{-\lambda(t + t_0)\}}{\exp\{-\lambda t_0\}} = \frac{\exp\{-\lambda t\}\exp\{-\lambda t_0\}}{\exp\{-\lambda t_0\}} = \exp\{-\lambda t\}$$

$$= P(X > t)$$

$P(\underline{X \leq t}) = \int_0^t \lambda e^{-\lambda t} dt = \boxed{1 - e^{-\lambda t}} = F(t)$

The Exponential Distribution/ RV

$X \sim Exp\left(\lambda = \frac{1}{1000}\right)$

Suppose a light bulb's lifetime is exponentially distributed with parameter λ. $= \frac{1}{1000}$

What is the probability that the lifetime of the light bulb lasts less than t $= 400$ hours?

$P(X < 400) = \int_0^{400} \lambda e^{-\lambda t} dt = \lambda e^{-\lambda t}\left(-\frac{1}{\lambda}\right) = -e^{-\lambda t}\Big|_0^{400} = -e^{-\lambda(400)} + e^{-\lambda(0)}$

$= 1 - e^{-\lambda(400)} = 0.329\ldots$

What is the probability that the lifetime of the light bulb lasts more than $t = 400$ hours?

$P(X > 400) = 1 - P(X < 400) = 1 - 0.329\ldots = 0.67\ldots$

The Exponential Distribution/ RV

$X \sim Exp\left(\frac{1}{1000}\right)$

Suppose a light bulb's lifetime is exponentially distributed with parameter λ: $= \frac{1}{1000}$

Now say you turn the light bulb on and then leave. You come back after t_0 hours to find it still on. What is the probability that the light bulb will last for *at least additional t* hours?

$t_0 = 400$, $t = 400$

$$P(\underbrace{X > 800}_{t+t_0} \mid \underbrace{X > 400}_{t_0}) = P(\underbrace{X > 400}_{t}) = 0.67\ldots$$

The Beta Distribution/RV

The exponential distribution has positive density over an infinite interval (so do many others, including the normal distribution, which we will learn about soon).

The *beta distribution* provides positive density only for X in an interval from 0 to 1.

The beta distribution is commonly used to model variation in the **proportion or percentage** of a quantity occurring in different samples.

$P(X<2)=1$

$\cdot P(X<0.5)=\int_0^{0.5} f(x;\alpha,\beta)\,dx$

The Beta Distribution/RV

Definition: A random variable *X* is said to have a *beta distribution* with parameters α and β (both positive), if the pdf of *X* is:

$$f(x;\alpha,\beta)=\underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{normalizing constant}}\underbrace{x^{\alpha-1}(1-x)^{\beta-1}}_{\text{defines shape}}$$

where:

$$\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-x}\,dx$$

In R:

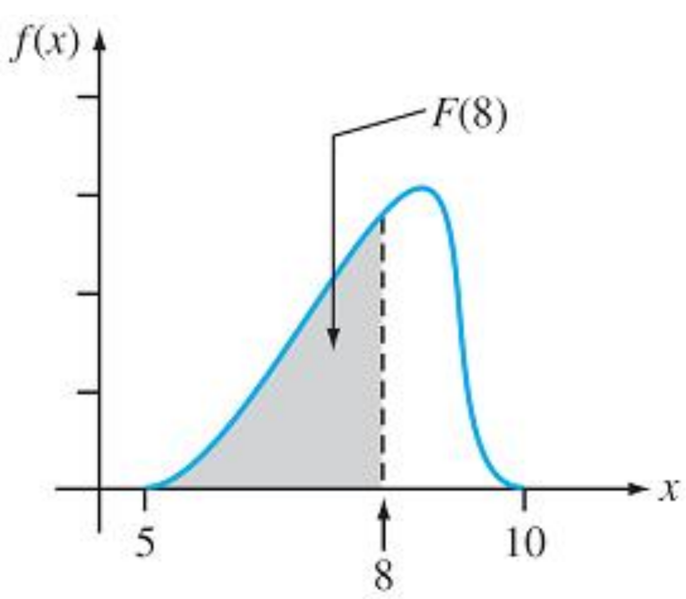
- rbeta( )
- pbeta( ) (cdf)
- dbeta( ) (pdf)

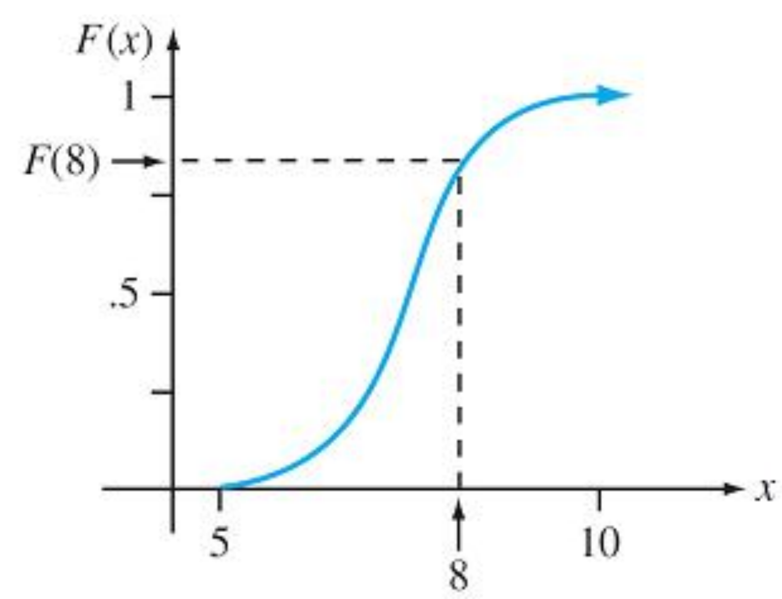
$F(x) = P(X \le x) = \sum_{t=0}^{x} \underline{P(X = t)}$

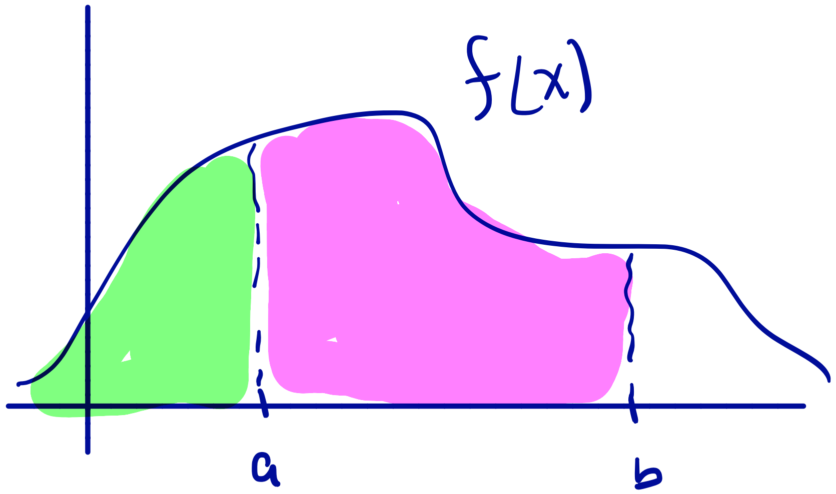
Cumulative Distribution Functions

Definition: The *cumulative distribution function* (*cdf*) is denoted with *F*(*x*). For a continuous *r.v. X* with pdf *f(x)*, *F(x)* is defined for every real number *x* by:

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







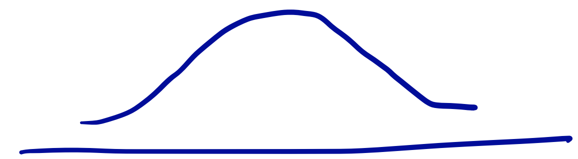
Example

$$\boxed{\int_{0.25}^{0.75} f(t)\,dt}$$

The distribution of the amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous rv *X* with pdf

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

1. What is the cdf of sales for any x? $F(x) = \int_0^x \frac{3}{2}(1-t^2)\,dt = \left.\frac{3}{2}t - \frac{3}{2}\frac{t^3}{3}\right|_0^x = \boxed{\frac{3}{2}x - \frac{x^3}{2}}$
2. Find the probability that X is less than .25?
   $P(X < 0.25) = F(0.25) = \dots = 0.37$
3. X is greater than .75?
   $P(X > 0.75) = 1 - P(X \le 0.75) = 1 - F(0.75) = 0.085\dots$
4. P(.25 < X < .75)?
   " $F(0.75) - F(0.25) = [1 - 0.085] - 0.37 = 0.544\dots$



Percentiles of a Distribution

Definition: The *median* of a continuous distribution is the 50th percentile of the distribution.

How can we express this in terms of *F(x)*?

$$\underset{\text{median}}{F(x)} = 0.5 \iff \int_{-\infty}^{x} f(t)\,dt = 0.5$$

Example

X is a r.v. such that:

$$\text{pdf} = f(x) = \frac{1}{2}(x+1), \qquad -1 < x < 1$$

1. Calculate F(x). $\int_{-1}^{x} \frac{1}{2}t + \frac{1}{2}\,dt = \frac{1}{4}t^2 + \frac{1}{2}t \Big|_{-1}^{x} = \frac{1}{4}x^2 + \frac{1}{2}x - \frac{1}{4} + \frac{1}{2} =$
   $= \frac{1}{4}x^2 + \frac{1}{2}x - \frac{1}{4} = F(x)$
2. What is the ~~0.375th~~ 37.5th percentile of this distribution?
   $\int_{-1}^{x} \frac{1}{2}(x+1)dx = 0.375 \iff \frac{1}{4}x^2 + \frac{1}{2}x - \frac{1}{4} = 0.375 \ldots$
3. What is the median of this distribution?
   $F(x) = 0.5$

Example

Suppose that we flip an unfair coin four times; the probability of flipping a head is 60%.

What is the median number of heads for this distribution?