CSCI 3022 Intro to Data Science
Continuous Random Variables

Opening: Recall: The Bernoulli A random variable whose only possible values are 0 or 1, with

pdf:

$$P(X=x)=f(x)=\begin{cases} p & x=1 \quad \text{called a ``sucess''} \\ (1-p) & x=0 \quad \text{called a ``failure''} \\ 0 & else \end{cases}$$

We often write $f(x)=p^x(1-p)^{1-x}$. Suppose we perform 3 independent Bernoulli experiments in a row, and then define $Y :=$ the number of successes in those 3 trials. What is $P(Y=2)$?

CSCI 3022 Intro to Data Science
Continuous Random Variables

Opening: Recall: The Bernoulli A random variable whose only possible values are 0 or 1, with

pdf:

$$P(X=x)=f(x)=\begin{cases} p & x=1 \quad \text{called a ``sucess''} \\ (1-p) & x=0 \quad \text{called a ``failure''} \\ 0 & else \end{cases}$$

We often write $f(x)=p^x(1-p)^{1-x}$. Suppose we perform 3 independent Bernoulli experiments in a row, and then define $Y :=$ the number of successes in those 3 trials. What is $P(Y=2)$?

$P(Y=2)=$ (# of ways you can choose 2 successes out of 3 trials) P (each one of those ways)

Opening Sol'n

Opening: The Bernoulli.

$$P(X = x)f(x) = p^x(1-p)^{1-x}$$

. Suppose we perform 3 independent Bernoulli experiments in a row, and then define $Y :=$ the number of successes in those 3 trials. What is $P(Y = 2)$?

Opening Sol'n

Opening: The Bernoulli.

$$P(X = x)f(x) = p^x(1-p)^{1-x}$$

. Suppose we perform 3 independent Bernoulli experiments in a row, and then define $Y :=$ the number of successes in those 3 trials. What is $P(Y = 2)$?

$P(Y = 2) = (\#$ of ways you can choose 2 successes out of 3 trials$)\, P\,($each one of those ways$)$

so: $P(Y = 2) = C(3, 2) \cdot P\,(\text{HHT}) = 3(p)^2(1-p)^1$ since each flip is independent.

Announcements and To-Dos

Announcements:

1. Another nb day this Friday.
2. No HW this week

Last time we learned:

1. about pdfs and cdfs; did some counting (3 major things: permutations, combinations, and the Binomial Theorem)

To do:

1. Check out the next set of notebooks!

Minute Forms:

Probability Distributions

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

If X is discrete, the pdf or probability mass function (pmf) f gives us $f(x) = P(X = x)$.

In the continuous case, the pdf instead gives probability to *intervals*.

Definition: *Cumulative Density Function*
For a discrete r.v. X with pdf $f(x) = P(X = x)$, the *cumulative density function*, denoted $F(x)$, is defined for every real number x to be the probability that the observed value of X will be at most x.
Mathematically: $F(x) = P(X \leq x)$

Trial and Error RVs

Bernoulli: A random variable whose only possible values are 0 or 1. Specified by a single parameter, the probability of a heads/"success" p! This gives the pdf:

$$P(X = x) = f(x) = p^x (1-p)^{1-x}$$

We denote the Bernoulli random variable X by $\underline{X \sim Bern(p)}$

Binomial: Let $X :=$ the number of successes of n trials of a Bern(p). Then:

$$P(X = i) = (\#\text{ of ways that } X = i) \cdot P(\text{of one such outcome})$$

$$P(X = i) = \binom{n}{i} \cdot P(n \text{ successes}) \cdot P(n - i \text{ failures}) = \binom{n}{i} p^i (1-p)^{(n-i)}$$

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}; \quad x \in \{0, 1, 2, \ldots, n\}$$

NOTATION: $\underline{X \sim bin(n, p)}$ for X a Binomial rv with success probability p and n trials.

The Binomial

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

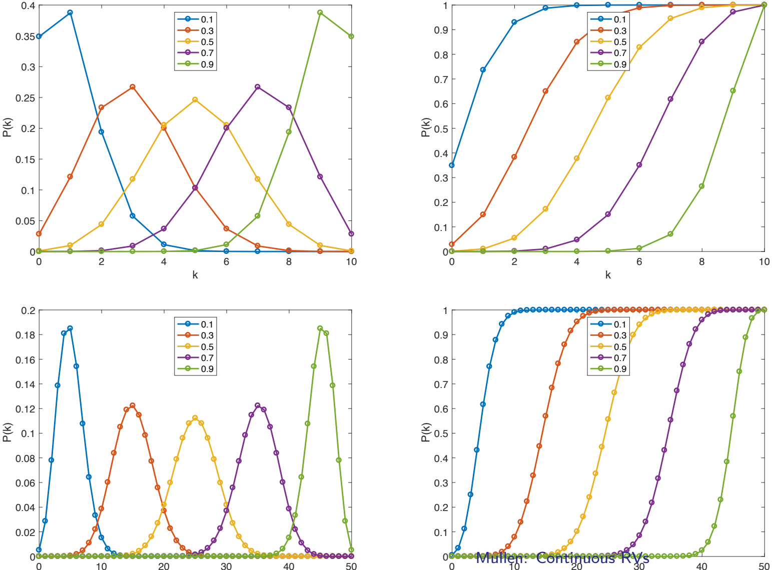
Important Assumptions:

1. Each trial must be *independent* of the previous experiment.
2. The probability of success must be *identical* for each trial.

The binomial is often defined and derived as the sum of n *independent, identically distributed* Bernoulli random variables.

In practice, any time we try to study a proportion on an underlying population, we gather a smaller sample where the observed proportion can often be thought of as a binomial random variable.

Binomial pdf and cdf. Top: $n = 10$; bottom: $n = 50$.



The Geometric

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.

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

(The per-donor probability checks are independent and identically distributed!)

The Geometric pdf

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

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 ____________ to indicate that X is a Geometric rv with success probability p.

The Geometric pdf

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

$$P(X = x) = P(\text{failed x-1 times}) \cdot P(\text{then success!})$$

$$P(X = x) = (1 - p)^{x-1} p; \quad x \in \{1, 2, 3, \ldots, \infty\}$$

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 $\underline{X \sim geom(p)}$ to indicate that X is a Geometric rv with success probability p.

The Geometric pdf

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

$$P(X = x) = P(\text{failed x-1 times}) \cdot P(\text{then success!})$$

$$P(X = x) = (1-p)^{x-1}p; \quad x \in \{1, 2, 3, \ldots, \infty\}$$

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 $\underline{X \sim geom(p)}$ to indicate that X is a Geometric rv with success probability p.
Important **note:** sometimes the geometric is counting the number of total *trials*; sometimes it's counting the number of *failures*. Know which one your software is doing!

The Negative Binomial

Motivating example:
A "successful toss" is defined to be the coin landing on heads. Let $X = \#$ of failures/tails before the *second* success/heads.

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

The Negative Binomial

Motivating example:
A "successful toss" is defined to be the coin landing on heads. Let $X = \#$ of failures/tails before the *second* success/heads.
Events in $X = 2$: $\{HTH, THH\}$
Events in $X = 3$: $\{HTTH, THTH, TTHH\}$
Events in $X = 4$: $\{HTTTH, THTTH, TTHTH, TTTHH\}$

How is this related to the geometric distribution? The binomial distribution?
It's like adding two geometrics.
The relationship to the binomial is a little harder, but if we know this random variables equals x, what do we know about trial $\#x$? The previous $x - 1$ trials?

The Negative Binomial

In general, let $X = \#$ of trials before the rth success. The pdf/pmf is:

NOTATION: We write ______________ to indicate that X is a Negative Binomial rv with success probability p and r successes until completion.

The Negative Binomial

In general, let $X = \#$ of trials before the rth success. The pdf/pmf is:

$$P(X = x) = (\#\text{ of ways that } X = x) \cdot P(\text{of one such outcome})$$

NOTATION: We write $\underline{X \sim NB(r, p)}$ to indicate that X is a Negative Binomial rv with success probability p and r successes until completion.

The Negative Binomial

In general, let $X = \#$ of trials before the rth success. The pdf/pmf is:

$$P(X = x) = (\#\text{ of ways that } X = x) \cdot P(\text{of one such outcome})$$

$$(\#\text{ of ways that } x-1 \text{ trials contain exactly } r-1 \text{ successes})$$

$$\cdot P(\text{r successes and } (x-1)-(r-1) \text{ failures}).$$

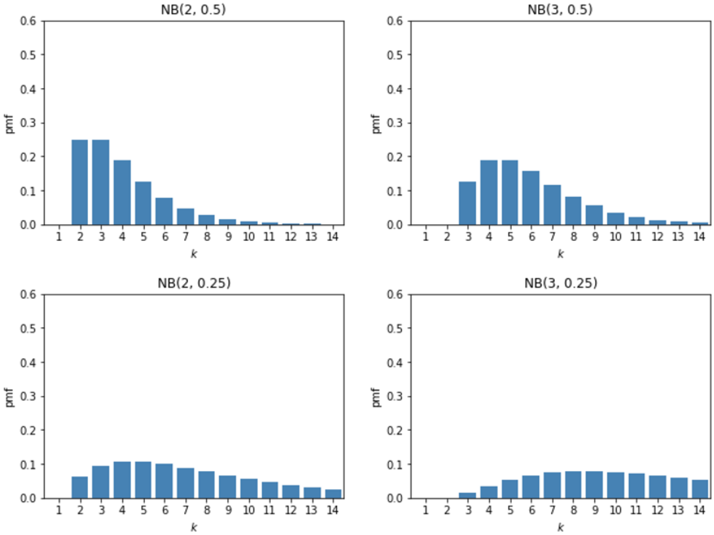
$$= \binom{x-1}{r-1} p^{r-1}(1-p)^{(x-1)-(r-1)} p$$

$$P(X = x) = \binom{x-1}{r-1} p^{r}(1-p)^{(x-r)}$$

for $x = \{r, r+1, r+2, \ldots \infty\}$.

NOTATION: We write $\underline{X \sim NB(r,p)}$ to indicate that X is a Negative Binomial rv with success probability p and r successes until completion.

NB pdfs


NB(2, 0.5)
NB(3, 0.5)
NB(2, 0.25)
NB(3, 0.25)
pmf
0.0
0.1
0.2
0.3
0.4
0.5
0.6
k
1
2
3
4
5
6
7
8
9
10
11
12
13
14

The Negative Binomial

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 exactly 15 people must be asked before 5 are found who agree to participate?

The Negative Binomial

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 exactly 15 people must be asked before 5 are found who agree to participate?

For $X \sim NB(5, .2)$, find $P(X = 15)$:

The Negative Binomial

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 exactly 15 people must be asked before 5 are found who agree to participate?

For $X \sim NB(5, .2)$, find $P(X = 15)$:

$$P(X = 15) = \binom{15-1}{5-1} .2^5 (.8)^{(15-5)}$$

The Poisson Distribution/RV

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

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

NOTATION: We write ____________ to indicate that X is a Poisson r.v. with parameter λ.

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

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

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

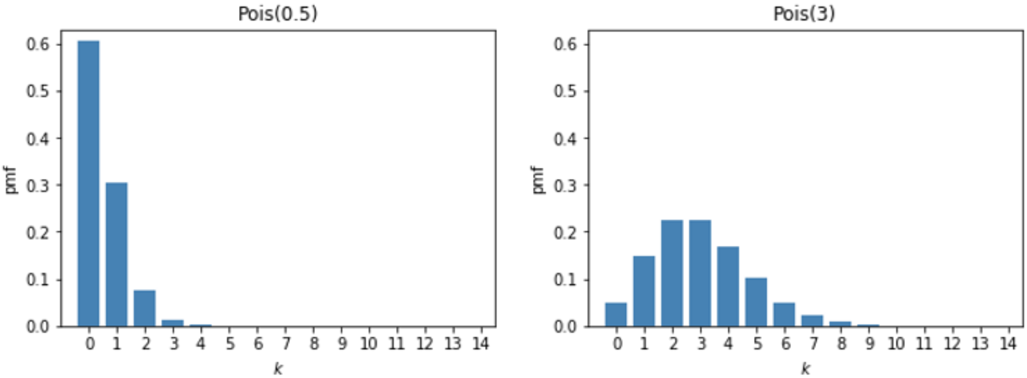
The Poisson Distribution/RV

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 $\lambda = 4.5$. What is the probability that the trap contains 5 mosquitoes?

The Poisson Distribution/RV

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 $\lambda = 4.5$. What is the probability that the trap contains 5 mosquitoes? $P(X = 5) =$

Poisson pdfs



Poisson and... binomial?

One way to generate the Poisson is to take limits of a binomial: suppose you get texts during class (☹) at a rate of 29 texts per hour. What is the probability that you get 29 texts in an hour? 12 texts in an hour? 107 texts in an hour?

λ is the *rate* of the Poisson.

Poisson and... binomial?

One way to generate the Poisson is to take limits of a binomial: suppose you get texts during class $\left(\ddot\frown\right)$ at a rate of 29 texts per hour. What is the probability that you get 29 texts in an hour? 12 texts in an hour? 107 texts in an hour?

λ is the *rate* of the Poisson.

Think about a Bernoulli that represents your friends asking "should I text...?" then flipping a coin with probability p. Then:

Poisson and... binomial?

One way to generate the Poisson is to take limits of a binomial: suppose you get texts during class $\left(\ddot{\frown}\right)$ at a rate of 29 texts per hour. What is the probability that you get 29 texts in an hour? 12 texts in an hour? 107 texts in an hour?

λ is the *rate* of the Poisson.

Think about a Bernoulli that represents your friends asking ”should I text...?” then flipping a coin with probability p. Then:

$\lambda = \frac{texts}{hour} \approx \frac{flips}{hour} \cdot \frac{texts}{flip} = np$ for the same n and p as a *binomial*.

Poisson and... binomial?

One way to generate the Poisson is to take limits of a binomial: suppose you get texts during class $(\ddot{\frown})$ at a rate of 29 texts per hour. What is the probability that you get 29 texts in an hour? 12 texts in an hour? 107 texts in an hour?

λ is the *rate* of the Poisson.

Think about a Bernoulli that represents your friends asking ”should I text...?” then flipping a coin with probability p. Then:

$\lambda = \frac{texts}{hour} \approx \frac{flips}{hour} \cdot \frac{texts}{flip} = np$ for the same n and p as a *binomial*.

...but n might vary a bit from hour to hour, so these are only equivalent *in the limit* (n large, p small)!

Continuous RVs

Many real-life random processes must be modeled by random variables that can take on continuous (non-discrete) values. Some example:

1. Peoples' heights: $X \in$
2. Final grades in a class: $X \in$
3. Time between people checking out at a store : $t \in$

Continuous RVs

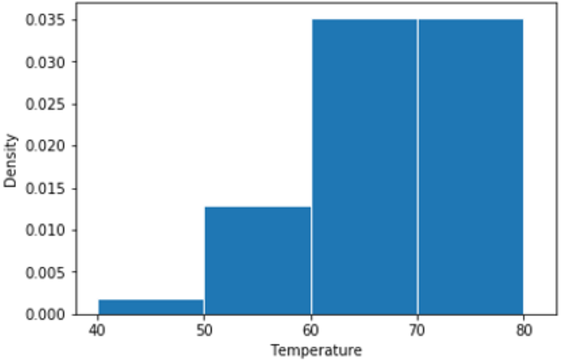
Many real-life random processes must be modeled by random variables that can take on continuous (non-discrete) values. Some example:

1. Peoples' heights: $X \in \{[0, 7.5ft]\}$
2. Final grades in a class: $X \in \{[0, 100]\}$
3. Time between people checking out at a store : $t \in \{[0, \infty]\}$

More precision!

Suppose your friend asks you what the high temperature will be today. They want to know the probability it will be between 70F and 80F, so they can wear shorts.

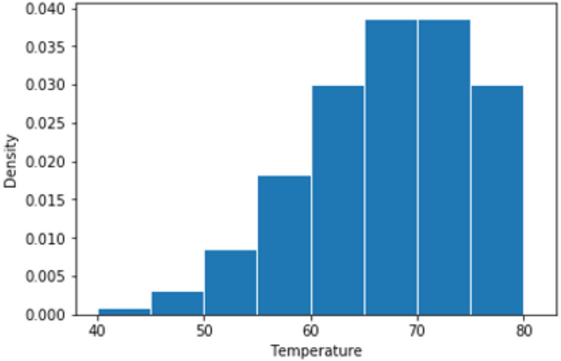
Probability:
Add up the share of outcomes between 70F and 80F!



More precision!

Suppose your friend asks you what the high temperature will be today. They want to know the probability it will be between 70F and 80F, so they can wear shorts.

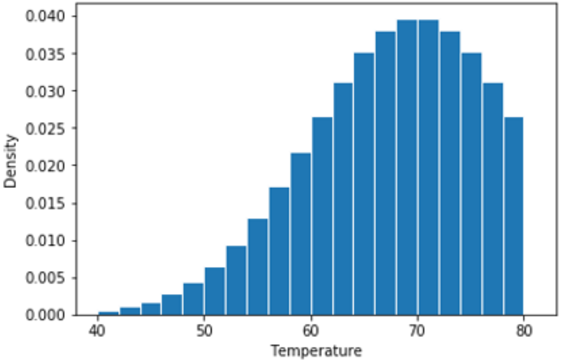
Probability:
Add up the share of outcomes between 70F and 80F!



More precision!

Suppose your friend asks you what the high temperature will be today. They want to know the probability it will be between 70F and 80F, so they can wear shorts.

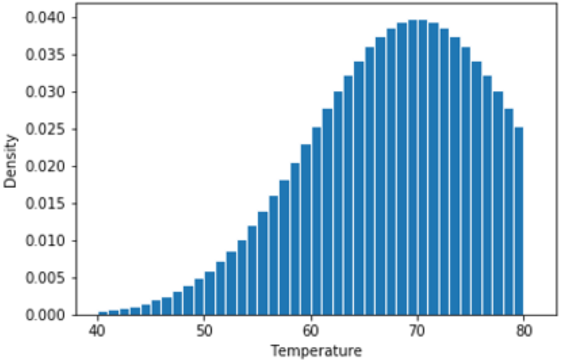
Probability:
Add up the share of outcomes between 70F and 80F!



More precision!

Suppose your friend asks you what the high temperature will be today. They want to know the probability it will be between 70F and 80F, so they can wear shorts.

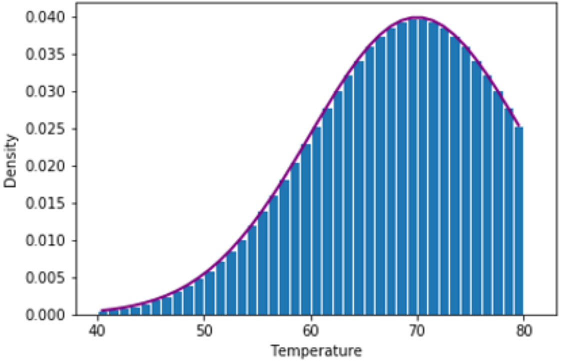
Probability:
Add up the share of outcomes between 70F and 80F!



More precision!

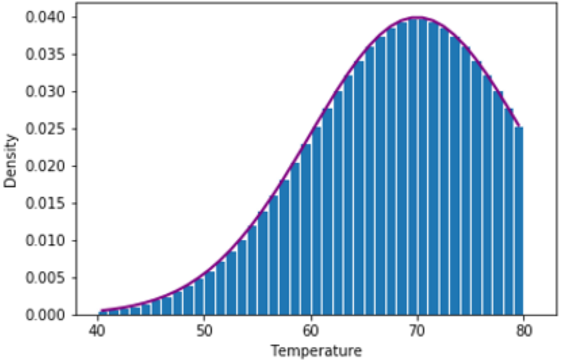
Suppose your friend asks you what the high temperature will be today. They want to know the probability it will be between 70F and 80F, so they can wear shorts.

Probability:
Add up the share of outcomes between 70F and 80F!



More precision!

Suppose your friend asks you what the high temperature will be today. They want to know the probability it will be between 70F and 80F, so they can wear shorts.



Probability:

Integrate up the share of outcomes between 70F and 80F!

Continuous Distributions

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:

$$f(x) = \begin{cases} \frac{1}{360} & 0 \leq X < 360 \\ 0 & else \end{cases}$$

Continuous Distributions

Example, cont'd:

$$f(x) = \begin{cases} \frac{1}{360} & 0 \leq X < 360 \\ 0 & else \end{cases}$$

Graphically, the pdf of X is:

Continuous Distributions

Example, cont'd: How can we show that:

1. the total area of the pdf of x is 1?

2. How do we calculate $P(90 \leq X \leq 180)$?

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

Continuous Distributions

Example, cont'd: How can we show that:

1. the total area of the pdf of x is 1?

$$\int_0^{360} f(x)\,dx = 1?$$

2. How do we calculate $P(90 \leq X \leq 180)$?

$$\int_{90}^{180} f(x)\,dx = \ldots?$$

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

$$P(X < 90 \textbf{ OR } X > 270) = \int_0^{90} f(x)\,dx + \int_{270}^{360} f(x)\,dx = \ldots?$$

Uniform Distribution

The previous problem was an example of the uniform distribution.

Definition: *Uniform Distribution*

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

NOTATION: We write __________ to indicate that X is a uniform rv with lower bound a and upper bound b.

Uniform Distribution

The previous problem was an example of the uniform distribution.

Definition: *Uniform Distribution*

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

$$f(x) = \frac{1}{b-a}; \qquad x \in [a, b]$$

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

Exponential Distribution

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

It can be thought of as a continuous analogue to the Poisson distribution, but instead of events-per-time, it measure time-per-events.

Examples:

Exponential Distribution

Definition: *Exponential Distribution*
A continuous rv X is said to have an *exponential distribution* with rate parameter λ if the pdf of X is:

NOTATION: We write ____________ to indicate that X is an exponential rv with rate λ.

Exponential Distribution

Definition: *Exponential Distribution*
A continuous rv X is said to have an *exponential distribution* with rate parameter λ if the pdf of X is:

$$f(x) = \lambda e^{-\lambda x}; \quad x \geq 0$$

NOTATION: We write $\underline{X \sim exp(\lambda)}$ to indicate that X is an exponential rv with rate λ.

Exponential Distribution

Example:

Suppose a light bulb's lifetime is exponentially distributed with parameter $\lambda = 1/1000$.

1. What are the units for λ?
2. What is the probability that the lifetime of the light bulb lasts less than 400 hours?
3. What is the probability that the lifetime of the light bulb lasts more than 5 hours?

Exponential Distribution

Example:

Suppose a light bulb's lifetime is exponentially distributed with parameter $\lambda = 1/1000$.

1. What are the units for λ?

   Same as Poisson: outcomes per time; so maybe burnouts per hour?

2. What is the probability that the lifetime of the light bulb lasts less than 400 hours?

$$P(X < 400) = \int_0^{400} \lambda e^{-\lambda x} = -e^{-\lambda x}|_0^{400} = 1 - e^{2/5}$$

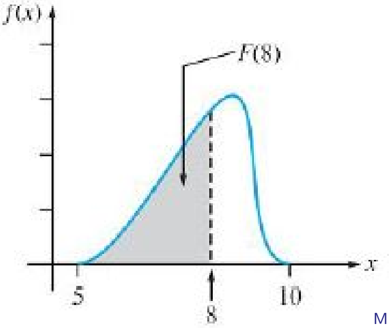
3. What is the probability that the lifetime of the light bulb lasts more than 5 hours?

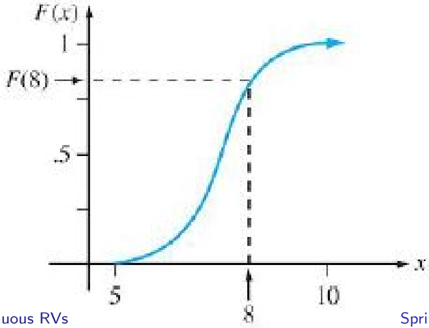
$$P(X > 5) = \int_5^{\infty} \lambda e^{-\lambda x} = -e^{-\lambda x}|_5^{\infty} = 0 - -e^{1/2000} \approx 1$$

Cumulative Density Function

Definition: *Cumulative Density Function*
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:

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





Continuous CDFs

Example:

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

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \leq X < 1 \\ 0 & else \end{cases}$$

1. What is the cdf of sales for any x?
2. Find the probability that X is less than .25?
3. X is greater than .75?
4. $P(.25 < X < .75)$?

Continuous CDFs

Example:

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

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \leq X < 1 \\ 0 & else \end{cases}$$

1. What is the cdf of sales for any x?
   $F(x) = P(X \leq x) = \int_0^x \frac{3}{2}(1-t^2)\,dt$
   $F(x) = \frac{3x}{2} - \frac{x^3}{2}$
2. Find the probability that X is less than .25? $F(.25)$
3. X is greater than .75? $1 - F(.75)$
4. $P(.25 < X < .75)$? $F(.75) - F(.25)$

Continuous CDFs

Wait, we've seen this before...

Recall: *The Fundamental Theorem of Calculus*.

Suppose F is an anti-derivative of f. Then:

1. $$\frac{d}{dx}\int_a^x f(t)\,dt = f(x);$$

   a.k.a.

   $$\frac{d}{dx}F(x) = f(x);$$

2. $$\int_a^b f(x)\,dx = F(B) - F(A).$$

Percentiles of a Distribution

Definition: The median $\tilde{x}$ of a continuous distribution is the 50th percentile or $.5$ quantile of the distribution.

How can we express this in terms of $f(x), F(x)$?

Notation:

Visually:

Percentiles of a Distribution

Definition: The median $\tilde{x}$ of a continuous distribution is the 50th percentile or $.5$ quantile of the distribution.

How can we express this in terms of $f(x), F(x)$?

Notation:

$\tilde{x}$ satisfies $F(\tilde{x}) = .5$, or

Visually:

$$.5 = \int_{\infty}^{\tilde{x}} f(x)\, dx$$

Daily Recap

Today we learned

1. Discrete and Continuous pdfs!

Moving forward:

- nb day Friday!
- No lecture Wednesday!

Next time in lecture:

- “Average values” of pdfs.