CSCI 3022 Intro to Data Science
Continuous Random Variables

Zach OH today!
3p - 6p.

Announcement and Reminders:

- HW3 Due tonight; HW4 posted tomorrow.
- Past Exams and solutions posted to Canvas! → Same format as yours.
- Exam is untimed, takehome. Designed to take around 2-3 hours, but you'll have most of a week. Due Mar 8, **likely** posted Mar 2. Think of it as an "all pen-and-paper" homework. Submission via Gradescope:
  1. You can TeX it yourself (template will be provided)
  2. You can take the given pdf and annotate on it
  3. You can print the pdf and write on it, then picture/scan your solutions
  4. You can work the exam on regular paper and picture/scan your work
- Tested content through lecture Mar 1 (Expectation & Variance)

Last Time...: the blocks of discrete probability

- Discrete uniform
"fair" n-sided die
Pick 1 card from deck, etc.

1. Bernoulli: *one* binary outcome experiment.
2. Binomial: binary outcome experiment success *count* in n tries.
3. Geometric: Total trials *until a success* of a binary outcome experiment.
4. Negative Binomial: Trials until *r* binary outcome experiment *successes*.
5. Poisson: *counting* outcomes with a fixed rate λ.

$\lambda = \frac{\text{count}}{\text{unit time}}$ (average).

Last Time...: the blocks of discrete probability

1. Bernoulli: *one* binary outcome experiment.
   $f(x) = p^x(1-p)^{1-x}$ input $x = \{0, 1\}$
2. Binomial: binary outcome experiment success *count* in n tries.
   $f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$
   $x = \{0, 1, 2, 3, \ldots, n\}$
3. Geometric: Total trials *until a success* of a binary outcome experiment.
   $f(x) = (1-p)^{x-1}p$
4. Negative Binomial: Trials until *r* binary outcome experiment *successes*.
   $f(x) = \binom{x-1}{r-1} p^r (1-p)^{(x-r)}$
5. Poisson: *counting* outcomes with a fixed rate λ.
   $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Last Time...: the blocks of discrete probability

The underlying pieces of discrete RVs:

1. The random variable X takes inputs/events in the (discrete) sample space Ω and maps them to a (discrete) finite or infinite set of probability values $a_1, a_2, a_3, \ldots$.
2. We find probabilities in the probability mass function or probability density function

$$f(x) = P(\underbrace{X = x}_{\text{specific } x \text{ values}}).$$

3. We can find cumulative probabilities or probability on ranges of outcomes in the cumulative density function

$$\boxed{F(x) = P(X \leq x)} = \sum_{X \leq x} f(x).$$

all outcomes
up to x.

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

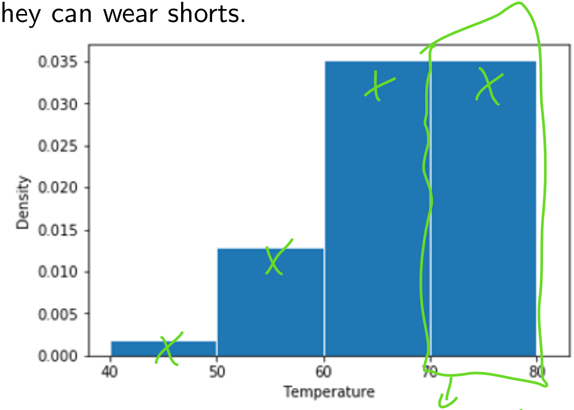
Many real-life random processes must be modeled by random variables that can take on continuous (non-discrete) values. Some example:

1. Peoples' heights: $X \in \{[\overset{3'}{\cancel{0}}, 7.5ft]\}$
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!

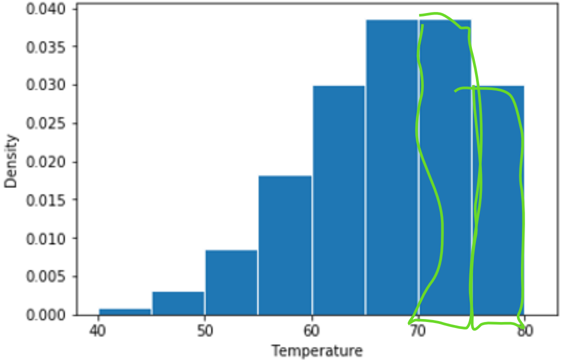


this rectangle is our probability

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.

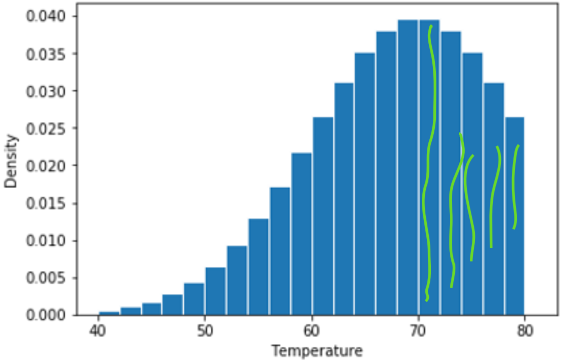
Probability:
Add up the share of outcomes between 70F and 80F!



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!

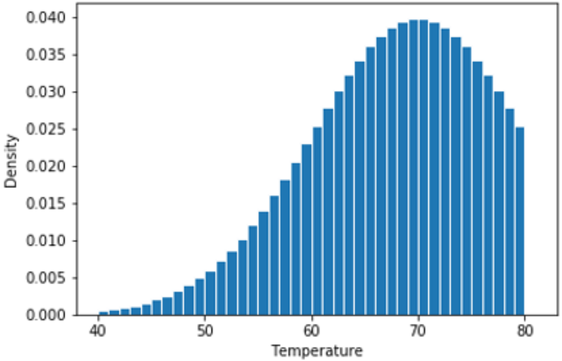


add these.

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.

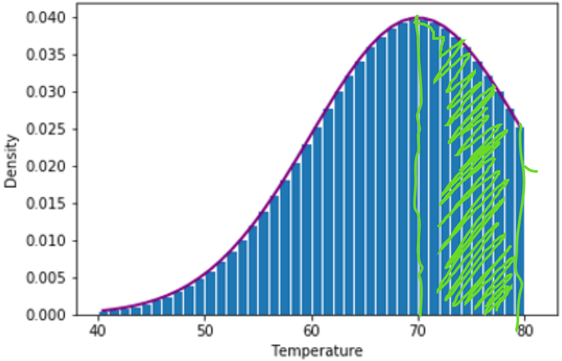
Probability:
Add up the share of outcomes between 70F and 80F!



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!



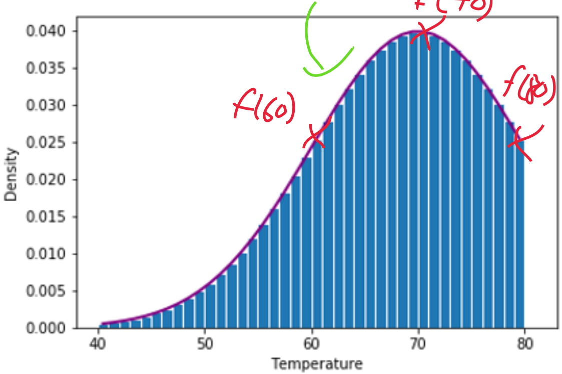
More precision!

there is a function for temperature

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: $f(x) = e^{-(x-70)^2}$

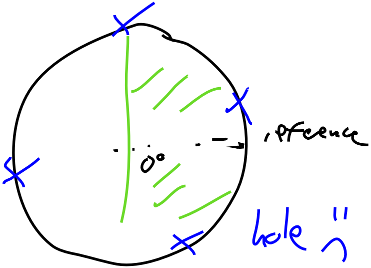
Integrate up the share of outcomes between 70F and 80F!



Continuous Distributions

Example:

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



all locations equally likely

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

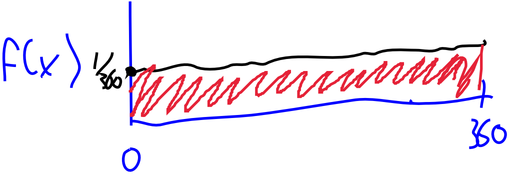
$$f(x) = \begin{cases} \boxed{\frac{1}{360}} & 0 \leq X < 360 \\ 0 & else \end{cases}$$

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:



note:

$$\boxed{P(0 \leq X < 360)} = P(\text{all outcomes}) = 1$$

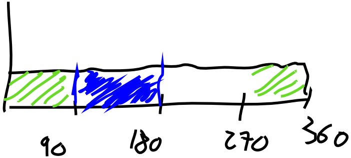
Area: length · width

$360 \cdot \frac{1}{360} = 1.$

Continuous Distributions

$f(x) = \frac{1}{360}$ for $x \in [0, 360)$

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



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

$$\int_0^{360} 1/360 \, dx = 1$$

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

length = 90
height = 1/360 $\Rightarrow$ Area = Probability = 1/4.

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

= 1/2

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 $P(90 \le X \le 180)$?

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

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

$\{[0,90)\} \cup \{(270,360\}$

exclusive

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

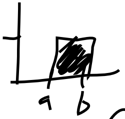
Uniform Distribution

rectangles: 1) know area = 1
2) choose min / left-hand side a
3) choose max / RHS. b

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:



width: b-a

height: 1/(b-a)

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le X \le b \\ 0 & \text{else} \end{cases}$$

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

$X \sim \text{unif}(a,b)$

$U(a,b)$

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.

Kinda like geometric!

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

rate λ units: $\frac{\text{count}}{\text{unit time}}$

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

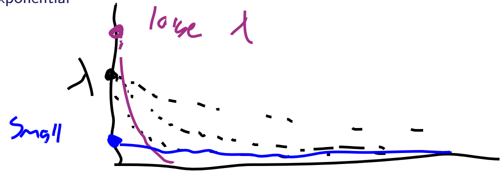
$$f(x) = \lambda \cdot e^{-\lambda x} \qquad x \in [0, \infty)$$

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

$X \sim \exp(\lambda)$.

Exponential Distribution

compare e^{-2x} to e^{-20x}



Definition: *Exponential Distribution*

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

at $x=0$, $f(0)=\lambda$

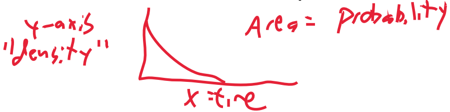
$$f(x) = \boxed{\lambda e^{-\lambda x}} \quad x \geq 0$$

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

units y · units x = "probability"

units y · "time" = probability

y = prob/"time"



Exponential Distribution

$X \sim \exp(1/1000)$

Example:

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

1. What are the units for λ?

$\frac{1 \text{ burnout}}{1000 \text{ hours}}$

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

$$P(X<400) = \int_0^{400} f(x)\,dx = \int_0^{400} \frac{1}{1000} \cdot e^{-\frac{x}{1000}}\,dx$$

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

integrate: chain rule $u = -\frac{x}{1000}$, $du = \frac{-dx}{1000}$ $\Rightarrow \int_{x=0}^{x=400} \frac{1}{1000} \cdot e^{u} \, (-1000\,du)$

$$= -\int_{x=0}^{x=400} e^{u}\,du = -e^{-x/1000}\Big|_0^{400}$$

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?

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

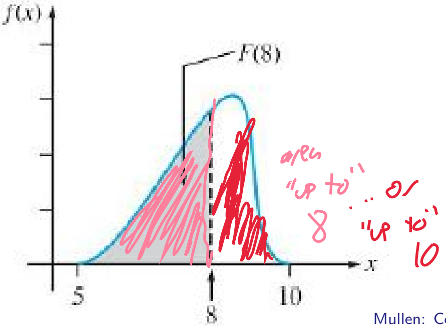
Cumulative Density Function

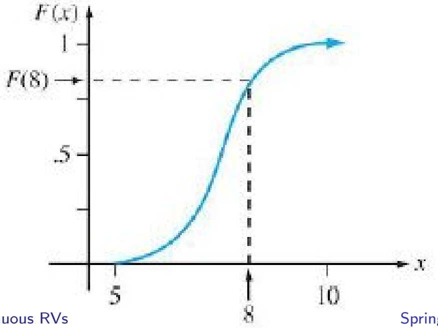
Definition: *Cumulative Density Function*
The *cumulative distribution function* (cdf) is denoted with $F(x)$. For a continuous r.v. X with pdf $f(x)$, $F(x)$ is defined for every real number x by:

Same as discrete!

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

notation: not x inside integral





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.

   $$\underbrace{\frac{d}{dx}F(x)}_{\frac{d}{dx}\text{ cdf}} = \underbrace{f(x)}_{=\text{ pdf.}};$$

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

   area of the pdf

$$\int_a^b f(x)dx = P(a \le X \le b) = F(b) - F(a)$$

we almost always need this integral!

Percentiles of a Distribution

Goals: find/describe properties of f & F; pdf & cdf.

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: $\int_a^b f(x)\,dx = P(a \le X \le b)$

we want "x" so that

$$\int_{start}^{x} f(t)\,dt = .5$$

left

right

$$= \int_{x}^{end} f(t)\,dt$$

Visually: mid point of the Area

what x-value cuts this function's area in half?

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

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

"start". 0 if exponential.

Visually:

Pops and Samples

Today marks the start of a large jump in how we approach data science problems:

1. We know about *sample statistics* like $\bar{X}$, s_X.
2. We've defined some *processes* that gives rise to distributions like the binomial, exponential, etc.
3. **Now:** we start bridging the gap! Given data and sample statistics, how do we estimate or infer properties of the underlying reality process? (parameters like p, λ).

   To do this, we need an understanding of centrality and dispersion of a process or density function might be.

Mean/Expected Value

Example:
Consider a university having 15,000 students and let X equal the number of courses for which a randomly selected student is registered.
The pdf of X is given to you as follows:

x	1	2	3	4	5	6	7
$f(x) = P(X = x)$	.01,	.03	.13	.25	.39	.17	.02

Students pay more money when enrolled in more courses, and so the university wants to know what the *average* number of courses students take per semester.

Mean/Expected Value

Definition: *Expected Value:*
For a discrete random variable X with pdf f(x), the *expected* value or *mean* value of X is denoted as E(X) and is calculated as:

Mean/Expected Value

Definition: *Expected Value:*
For a discrete random variable X with pdf f(x), the *expected* value or *mean* value of X is denoted as E(X) and is calculated as:

$$E[X] = \sum_{x \in \Omega} x \cdot P(X = x)$$

Mean/Expected Value

Example:, cont'd:

The pdf of X is given to you as follows:

x	1	2	3	4	5	6	7
$f(x) = P(X = x)$	.01,	.03	.13	.25	.39	.17	.02

What is $E[X]$?

Mean/Expected Value

Example:, cont'd:
The pdf of X is given to you as follows:

x	1	2	3	4	5	6	7
$f(x) = P(X = x)$	.01,	.03	.13	.25	.39	.17	.02

What is $E[X]$?

$$E[X] = \sum_{x \in \Omega} x \cdot P(X = x) = 1 \cdot .01 + 2 \cdot .03 + 3 \cdot .13 + 4 \cdot .25 + 5 \cdot .39 + 6 \cdot .17 + 7 \cdot .02$$

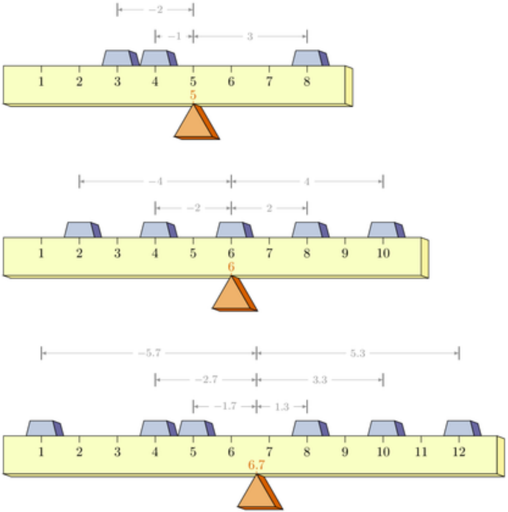
$$E[X] = 4.57$$

Interpreting Expected Value: Relative Frequency

One way to interpret expected value of a discrete distribution (especially on a finite support) is the sample mean if we managed to observe observations that *exactly* mirror the probability mass function.

In the preceding example, the pmf was given at 7 values of X with a precision up to 1%. In this case, if we had exactly 100 students and their proportions *observed* exactly mirrored the probabilities given in the example, the sample mean would be identical to the population mean.

Interpreting Expected Value



- The "center of mass" of a set of point masses
- Each mass exerts an "$r \times f$" force on the balancing point.
- Same idea holds in continuous space: we're looking for a centroid of an object.

http://www.texample.net/media/tikz/examples/TEX/balance.tex

Daily Recap

Today we learned

1. Continuous pdfs!

Moving forward:

- nb day Friday!

Next time in lecture:

- More on “average values” of pdfs, plus calculating variances from them!