CSCI 3022

intro to data science with probability & statistics

Lecture 10
February 16, 2018

Expected Values

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

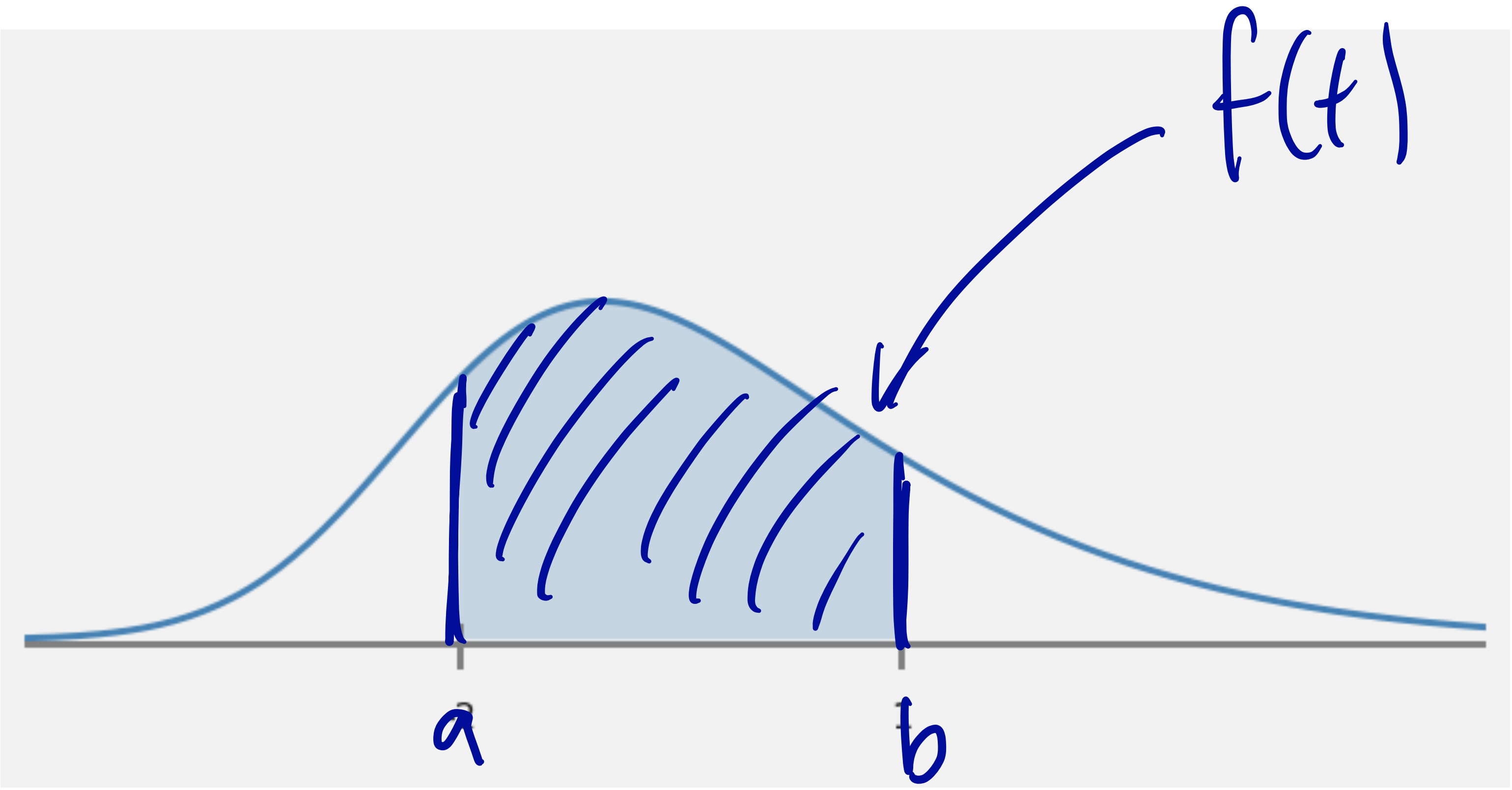
Dan Larremore

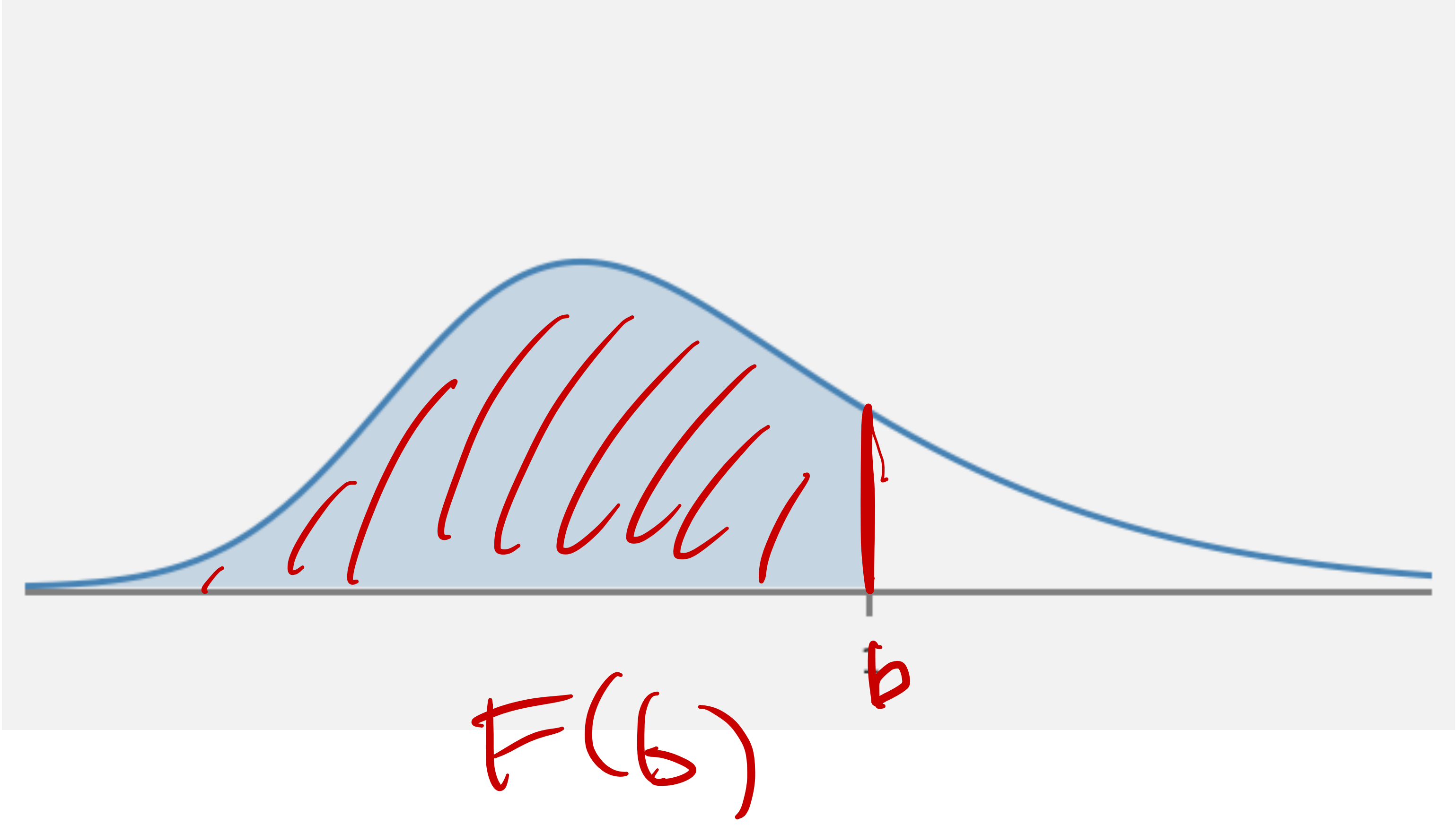
Last time on CSCI 3022:

PDF $f(t)$

- **Continuous random variables**:

$$P(a \leq X \leq b) = \int_a^b f(t)dt = F(b) - F(a)$$



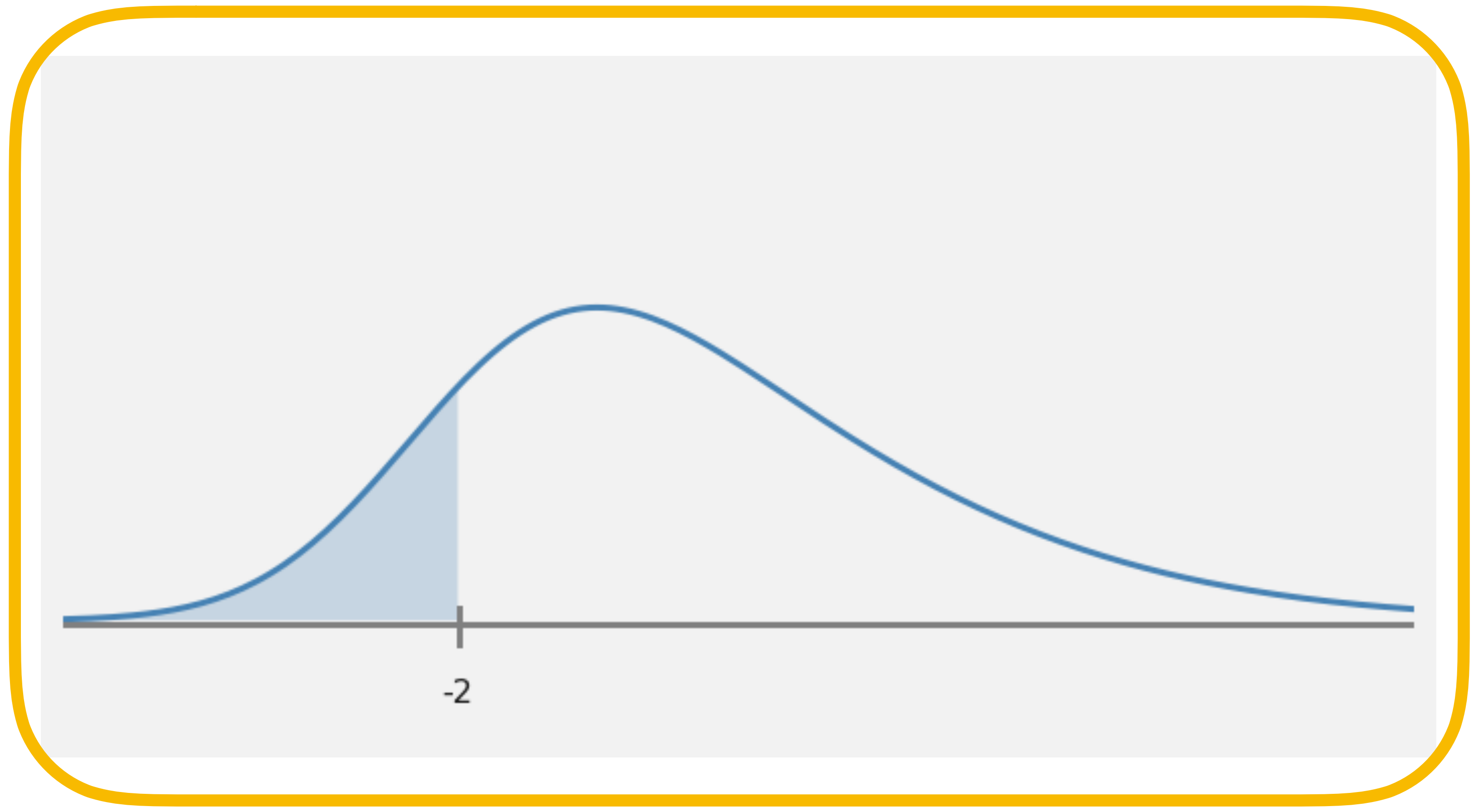


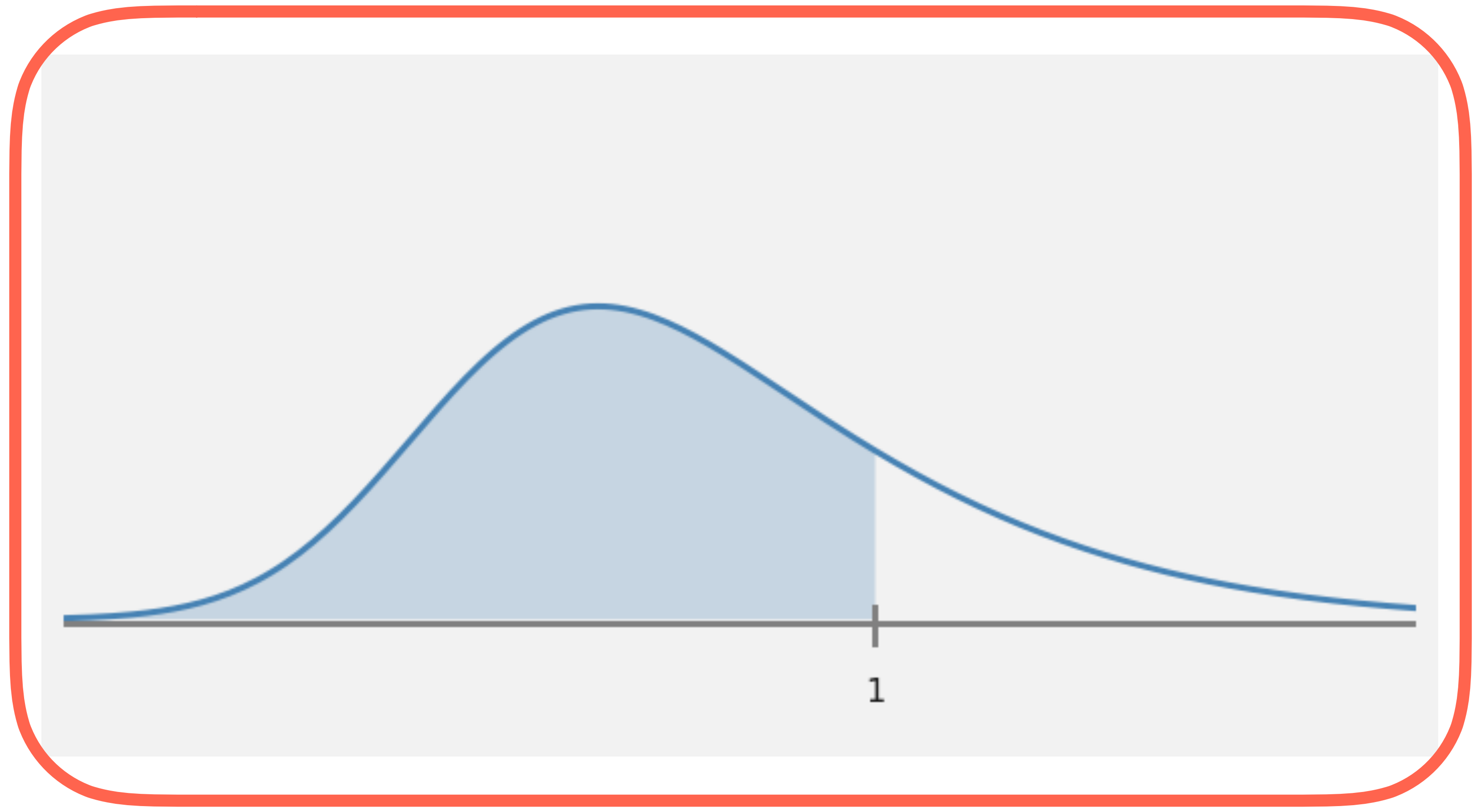
- New distributions! Uniform, Exponential, Normal.

Last time on CSCI 3022:

- **Continuous random variables**:

$$P(a \leq X \leq b) = \boxed{\int_a^b f(t)dt} = \boxed{F(b)} - \boxed{F(a)}$$





- New distributions! Uniform, Exponential, Normal.

Homework Planning

weighted average = expected value

- Suppose *hypothetically* that I write the homework questions as either: easy (takes 10 mins), medium (30 mins), or hard (60 mins).
- The probability that each question is easy, medium, or hard, is: 0.4, 0.35, 0.25, respectively.
- If a homework consists of 5 questions, what's the average time it takes to do the homework? answer ~ minutes

weighted average problem per

\mu

$$\underbrace{0.4}_{Pr(easy)} \cdot \underbrace{10}_{\#mins\ easy} + \underbrace{0.35}_{Pr(med)} \cdot \underbrace{30}_{\#mins\ med} + \underbrace{0.25}_{Pr(hard)} \cdot \underbrace{60}_{\#mins\ hard} = \mu$$

$$\boxed{5\mu}$$

Expected Value

$nx^{n-1} = \frac{d}{dx} x^n$

- **Definition**: The *expectation* or *expected value* of a discrete random variable *X* that takes the values $a_1, a_2, \ldots$ and with PMF *p* is given by:

minutes

$$\mathrm{E}[X] = \sum_i a_i P(X = a_i) = \sum_i a_i \underbrace{p(a_i)}_{\text{Pr of getting that kind of problem (Weight)}}$$

geometric

$p(k) = p(1-p)^{k-1}$

- **Exercise**: What is the expected value of the geometric distribution?

$$E[X] = \sum_{k=1}^{\infty} \underbrace{k}_{a_i} \underbrace{p(1-p)^{k-1}}_{p(a_i)}$$

$n = k-1$
$k = n+1$

$$= p \sum_{k=1}^{\infty} k (1-p)^{k-1}$$

$$= p \sum_{n=0}^{\infty} (n+1)(1-p)^n$$

$$p\left[\sum_{n=0}^{\infty} (1-p)^n + \sum_{n=0}^{\infty} n(1-p)^n\right]$$

$$p\left[\frac{1}{p} + \sum_{n=0}^{\infty} n(1-p)^n\right]$$

$$1 + p \sum_{n=0}^{\infty} n \underbrace{(1-p)^n}_{(1-p)^{n-1}(1-p)}$$

$$1 + p(1-p) \sum_{n=0}^{\infty} n(1-p)^{n-1}$$

$$= 1 + p(1-p) \frac{d}{dp}\left(-\sum_{n=0}^{\infty} (1-p)^n\right)$$

$$= 1 + p(1-p) \frac{d}{dp}\left(-\frac{1}{p}\right)$$

$$= 1 + \frac{\cancel{p}(1-p)}{p^{\cancel{2}}}$$

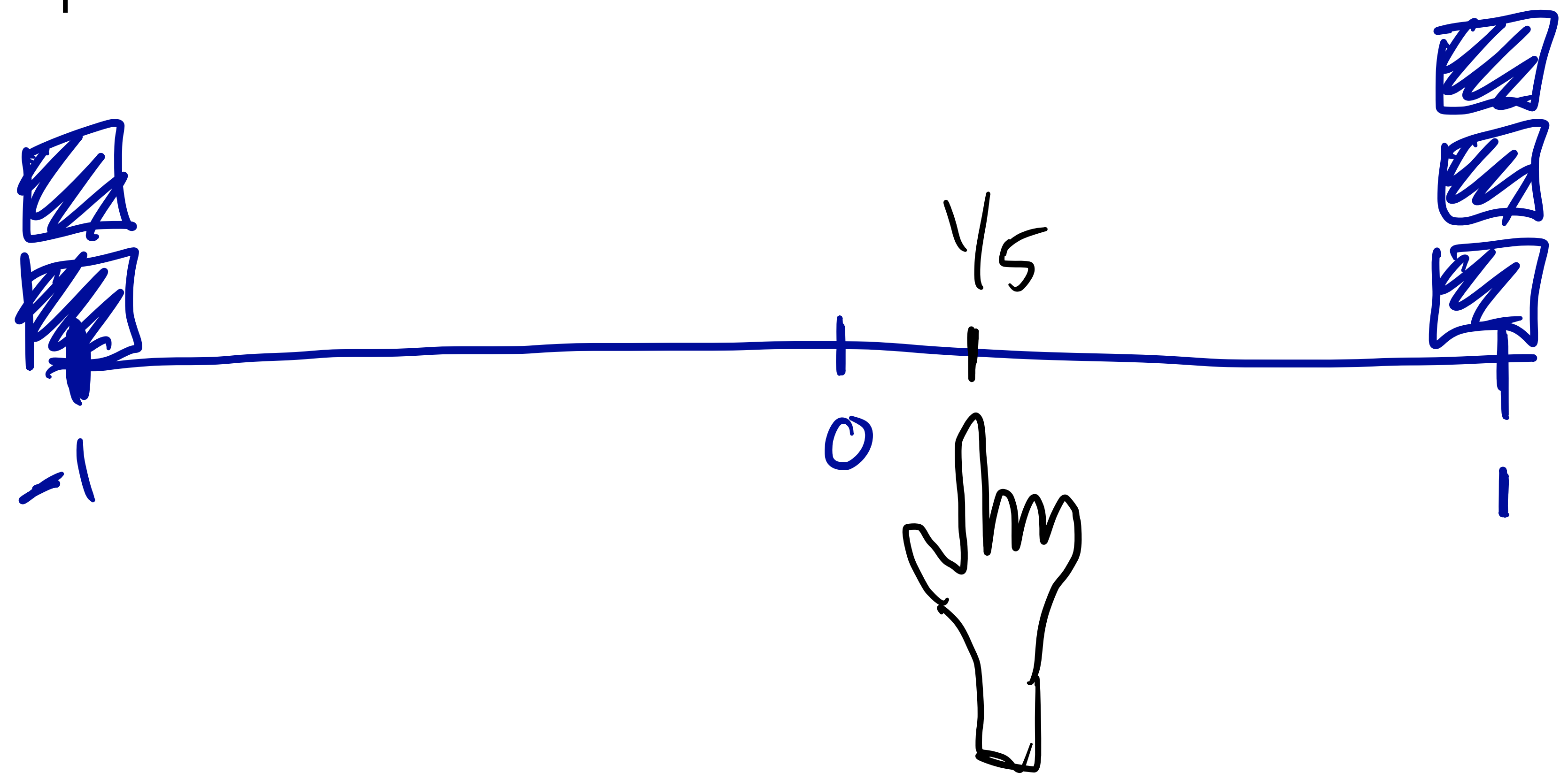
$$= 1 + \frac{1}{p} - \frac{p}{p} = \boxed{\frac{1}{p}}$$

:)

Expected Value: center of gravity

- **Note**: the expected value is the *center of gravity*.

- **Example**: suppose I stack 2 boxes at x=-1 and 3 boxes at position x=1. What is the expected value of this distribution of boxes?



$P(-1) = 2/5$

$P(1) = 3/5$

$E[X] = \sum_i a_i \, p(a_i)$

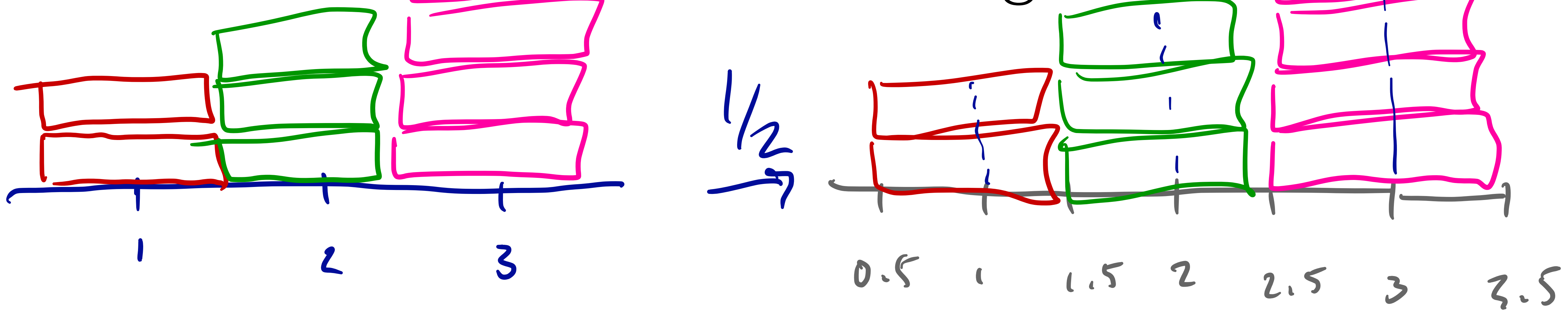
$$E[X] = a_{-1}\, p(-1) + a_1\, p(1)$$

$$= (-1)\left(\frac{2}{5}\right) + (1)\left(\frac{3}{5}\right) = -\frac{2}{5} + \frac{3}{5} = \boxed{\frac{1}{5}}$$

What about continuous RVs?

- Back to box stacking. What if I have 2, 3, and 5 *physical* IRL boxes at positions 1, 2, and 3, respectively.

- The masses are not concentrated at a point. We can change our discretization, subdividing, and subdividing…



from PMF to the limit: PDF

$\frac{2}{10}$ $\frac{3}{10}$ $\frac{5}{10}$

$\frac{2}{20}$ $\frac{2}{20}$ $\frac{3}{20}$ $\frac{3}{20}$ $\frac{5}{20}$ $\frac{5}{20}$

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

$$\sum a_i \, p(a_i)$$

What about continuous RVs?

- Back to box stacking. What if I have 2, 3, and 5 *physical* IRL boxes at positions 1, 2, and 3, respectively.
- The masses are not concentrated at a point. We can change our discretization, subdividing, and subdividing…
- **Definition**: The *expectation* or *expected value* of a continuous random variable X with PDF f is the number:

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

\infty

Expected value: average, c. of. g

- The expected value E[X] is also the average of a large number of draws of the random variable X.

- Even in the continuous case, E[X] is the center of gravity.

- **Example**: What is the expectation of an exponential distribution?

$X \sim \exp(\lambda)$

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

def'n of exponential

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

$$= \int_{0}^{\infty} x \lambda e^{-\lambda x}\,dx$$

Integration by parts! ^_^;"

$$= \boxed{\frac{1}{\lambda}}$$

recall geometric

$E[X] = \frac{1}{p}$

Expected value of a normal

- Let $X \sim N(\mu, \sigma^2)$
- Then: $\mathrm{E}[X] = \int_{-\infty}^{\infty} \underbrace{x}_{x} \underbrace{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}_{f(x)} dx = \mu$

(1) change of variables $z = x - \mu$

$dz = dx$

$x = z + \mu$

$$\underbrace{\int_{-\infty}^{\infty} z \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2} dz}_{0} + \underbrace{\int_{-\infty}^{\infty} \mu \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2} dz}_{\mu \cdot \text{normal}}$$

bc odd

$+ \mu \cdot 1$

$= \boxed{\mu}$

Change of variable trick

$g(x) = x$

$g(x) = x^2$ $\quad g(x) = \sin(x)$

- Let X and Y be random variables and let $g : \mathcal{R} \rightarrow \mathcal{R}$ be a function.

- For X discrete and Y continuous, then:

$$\mathrm{E}[g(X)] = \sum_i g(a_i) f(a_i) \qquad\qquad \mathrm{E}[g(\cancel{X}Y)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

If $g(x) = x$

$$E[g(\cancel{X})] = \sum_i a_i f(a_i) = E[X]$$

$$\shortparallel$$

$$E[X]$$

If $g(x) = x^2$

$$E[g(x)] = \sum_i (a_i)^2 f(a_i)$$

$$\shortparallel$$

$$E[X^2]$$

Change of variable trick

- Let X and Y be random variables and let $g : \mathcal{R} \rightarrow \mathcal{R}$ be a function.

- For X discrete and Y continuous, then:

$$E[g(X)] = \sum_i g(a_i) f(a_i)$$

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

- What happens if g(x) = x? ✓

- What happens if g(x) = rX +s? ✓

$$\boxed{E[rX+s] = rE[X]+s}$$

$$E[rX+s] = \sum_i (ra_i + s) f(a_i) = \sum_i r a_i f(a_i) + \sum_i s f(a_i)$$

$$= r\sum_i a_i f(a_i) + s\boxed{\sum_i f(a_i)} = rE[X] + s$$