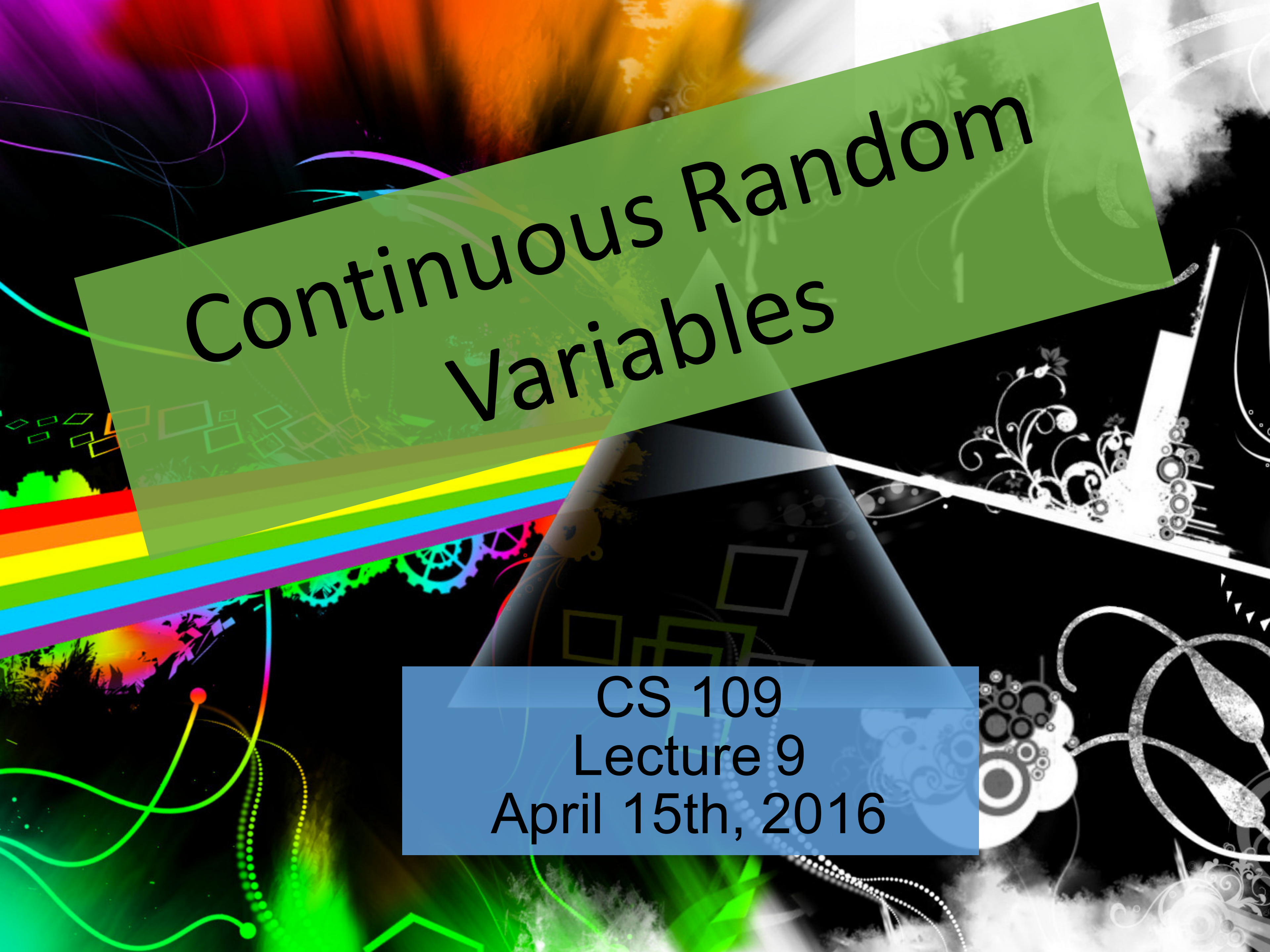


Continuous Random Variables

CS 109
Lecture 9
April 15th, 2016

Review

Discrete Distributions

Bernoulli:

- indicator of coin flip X ~ Ber(*p*)

Binomial:

- # successes in *n* coin flips X ~ Bin(*n*, *p*)

Poisson:

- # successes in *n* coin flips X ~ Poi(λ)

Geometric:

- # coin flips until success X ~ Geo(*p*)

Negative Binomial:

- # trials until *r* successes X ~ NegBin(*r*, *p*)

Hyper Geometric:

- # white balls drawn without replacement from urn with *N* balls, *m* are white: X ~ HypG(*n*, *N*, *m*)

Balls, Urns and the Supreme Court

Supreme Court case: Berghuis v. Smith

If a group is underrepresented in a jury pool, how do you tell?

- Article by Erin Miller – Friday, January 22, 2010
- Thanks to (former CS109er) Josh Falk for this article

Justice Breyer [Stanford Alum] opened the questioning by invoking the binomial theorem. He hypothesized a scenario involving **"an urn with a thousand balls, and sixty are blue, and nine hundred forty are purple, and then you select them at random… twelve at a time."** According to Justice Breyer and the binomial theorem, if the purple balls were under represented jurors then **"you would expect… something like <u>a third to a half</u> of juries would have at least one minority person"** on them.

Justin Breyer Meets CS109

- Should model this combinatorially (X ~ HypGeo)
    - Ball draws not independent trials (balls not replaced)
- Exact solution:

$$P(\text{draw 12 purple balls}) = \binom{940}{12} \Big/ \binom{1000}{12} \approx 0.4739$$

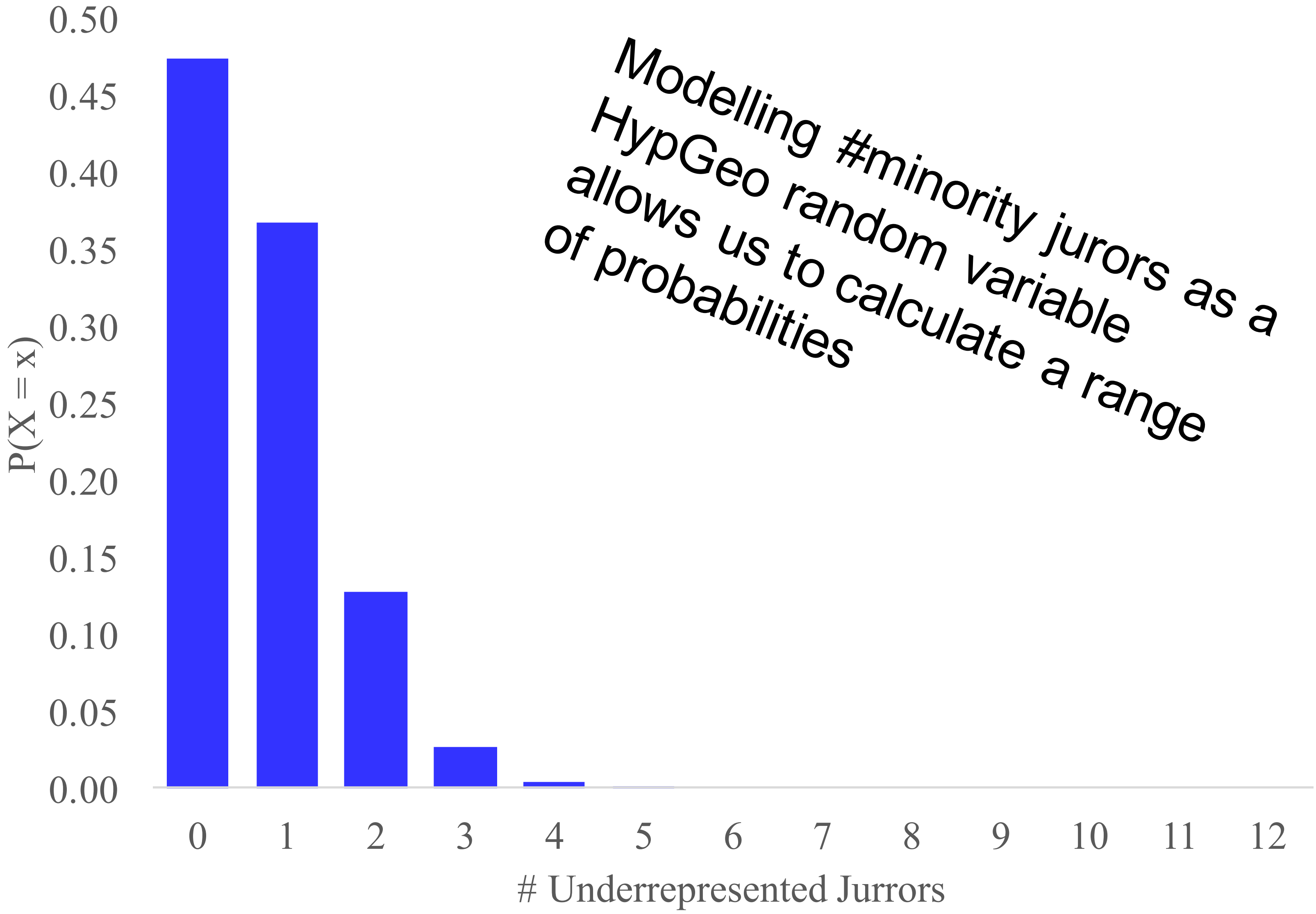
$$P(\text{draw} \geq 1 \text{ blue ball}) = 1 - P(\text{draw 12 purple}) \approx 0.5261$$

- Approximation using Binomial distribution
    - Assume P(blue ball) constant for every draw = 60/1000
    - X = # blue balls drawn. X ~ Bin(12, 60/1000 = 0.06)
    - $P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0.4759 = 0.5240$

In Breyer's description, should actually expect just <u>over half</u> of juries to have at least one black person on them

Demo

Underrepresented Juror PMF



Modelling #minority jurors as a HypGeo random variable allows us to calculate a range of probabilities

Endangered Species

- Determine *N* = how many of some species remain
    - Randomly tag *m* of species (e.g., with white paint)
    - Allow animals to mix randomly (assuming no breeding)
    - Later, randomly observe another *n* of the species
    - X = number of tagged animals in observed group of *n*
    - X ~ HypG(*n*, *N*, *m*)
- "Maximum Likelihood" estimate
    - Set *N* to be value that maximizes: $P(X = i) = \dfrac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$

      for the value *i* of X that you observed → $\hat{N}$ = *mn*/*i*
- Calculated by assuming: *i* = E[X] = *nm*/*N*

End Review

From Discrete to Continuous

- So far, all random variables we saw were *discrete*
  - Have finite or countably infinite values (e.g., integers)
  - Usually, values are binary or represent a count
- Now it's time for *continuous* random variables
  - Have (uncountably) infinite values (e.g., real numbers)
  - Usually represent measurements (arbitrary precision)
    - Height (centimeters), Weight (lbs.), Time (seconds), etc.
- Difference between how *many* and how *much*
- Generally, it means replace $\sum_{x=a}^{b} f(x)$ with $\int_{a}^{b} f(x)dx$

Integrals

*loving, not scary

Continuous Random Variables

- X is a **Continuous Random Variable** if there is function $f(x) \geq 0$ for $-\infty \leq x \leq \infty$, such that:

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

- f is a Probability Density Function (PDF) if:

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

Probability Density Function

- Say f is a **Probability Density Function** (PDF)

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

  - $f(x)$ is **not** a probability, it is probability/units of X
  - Not meaningful without some subinterval over X

$$P(X = a) = \int_{a}^{a} f(x)dx = 0$$

  - Contrast with Probability Mass Function (PMF) in discrete case: $p(a) = P(X = a)$

    where $\sum_{i=1}^{\infty} p(x_i) = 1$ for X taking on values x_1, x_2, x_3, …

Cumulative Distribution Function

- For a continuous random variable X, the **Cumulative Distribution Function** (CDF) is:

$$F(a) = P(X < a) = P(X \le a) = \int_{-\infty}^{a} f(x)dx$$

- Density f is derivative of CDF F: $f(a) = \frac{d}{da}F(a)$
- For continuous f and small ε :

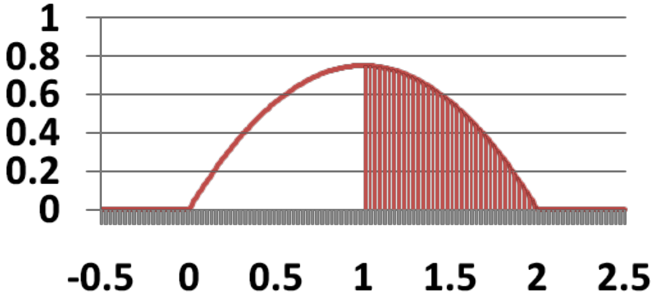
$$P(a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}) = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)dx \approx \varepsilon f(a)$$

  - So, ratio of probabilities can still be meaningful:
    - P(X = 1)/P(X = 2) ≈ (ε *f*(1))/(ε *f*(2)) = *f*(1)/*f*(2)

Simple Example

- X is continuous random variable (CRV) with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$



  - What is C?

$$\int_0^2 C(4x - 2x^2)dx = 1 \quad \Rightarrow \quad C\left(2x^2 - \frac{2x^3}{3}\right)\Bigg|_0^2 = 1$$

$$C\left(\left(8 - \frac{16}{3}\right) - 0\right) = 1 \quad \Rightarrow \quad C\frac{8}{3} = 1 \quad \Rightarrow \quad C = \frac{3}{8}$$

  - What is P(X > 1)?

$$\int_1^\infty f(x)dx = \int_1^2 \frac{3}{8}(4x - 2x^2)dx = \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right)\Bigg|_1^2 = \frac{3}{8}\left[\left(8 - \frac{16}{3}\right) - \left(2 - \frac{2}{3}\right)\right] = \frac{1}{2}$$

Disk Crashes

- X = days of use before your disk crashes

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

  - First, determine λ to have actual PDF
    - Good integral to know: $\int e^u \, du = e^u$

$$1 = \int \lambda e^{-x/100} dx = -100\lambda \int \frac{-1}{100} e^{-x/100} dx = -100\lambda e^{-x/100} \Big|_0^\infty = 100\lambda \Rightarrow \lambda = \frac{1}{100}$$

  - What is P(50 < X < 150)?

$$F(150) - F(50) = \int_{50}^{150} \frac{1}{100} e^{-x/100} dx = -e^{-x/100} \Big|_{50}^{150} = -e^{-3/2} + e^{-1/2} \approx 0.383$$

  - What is P(X < 10)?

$$F(10) = \int_0^{10} \frac{1}{100} e^{-x/100} dx = -e^{-x/100} \Big|_0^{10} = -e^{-1/10} + 1 \approx 0.095$$

Expectation and Variance

For discrete RV X:

$$E[X] = \sum_{x} x\, p(x)$$

$$E[g(X)] = \sum_{x} g(x)\, p(x)$$

$$E[X^n] = \sum_{x} x^n\, p(x)$$

For continuous RV X:

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

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

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

For both discrete and continuous RVs:

$$E[aX + b] = aE[X] + b$$

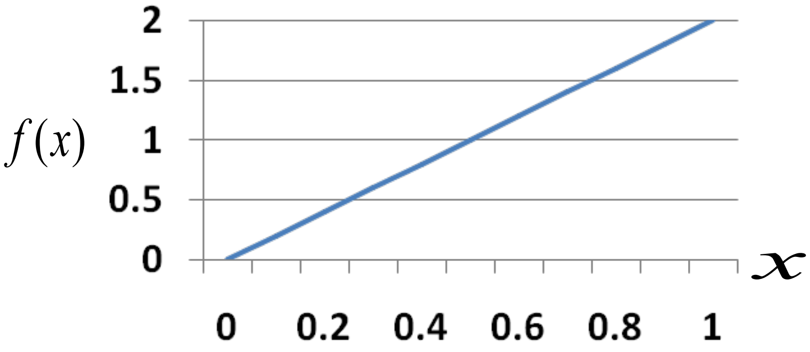
$$\mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$$

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$$

Linearly Increasing Density

- X is a continuous random variable with PDF:

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



  - What is E[X]?

$$E[X] = \int_{-\infty}^{\infty} x\, f(x) dx = \int_0^1 2x^2\, dx = \frac{2}{3} x^3 \Big|_0^1 = \frac{2}{3}$$

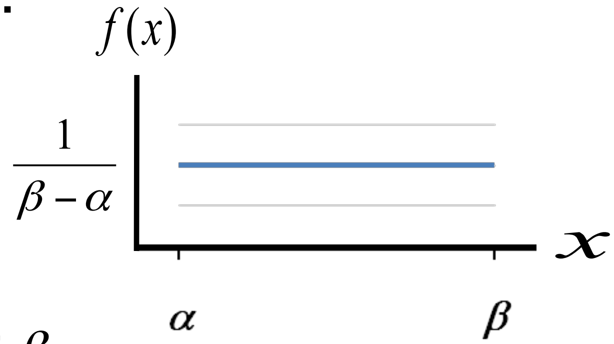
  - What is Var(X)?

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 2x^3\, dx = \frac{1}{2} x^4 \Big|_0^1 = \frac{1}{2}$$

$$Var(X) = E[X^2] - (E[X])^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

Uniform Random Variable

- X is a **Uniform Random Variable**: $X \sim \text{Uni}(\alpha, \beta)$
  - Probability Density Function (PDF):



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

    - Sometimes defined over range $\alpha < x < \beta$
  - $P(a \le x \le b) = \int_a^b f(x)dx = \frac{b-a}{\beta-\alpha}$ (for $\alpha \le a \le b \le \beta$)
  - $E[X] = \int_{-\infty}^{\infty} x f(x)dx = \int_{\alpha}^{\beta} \frac{x}{\beta-\alpha}dx = \left.\frac{x^2}{2(\beta-\alpha)}\right|_{\alpha}^{\beta} = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{\alpha+\beta}{2}$
  - $Var(X) = \frac{(\beta-\alpha)^2}{12}$

Fun with the Uniform Distribution

- X ~ Uni(0, 20)

$$f(x) = \begin{cases} \frac{1}{20} & 0 \le x \le 20 \\ 0 & \text{otherwise} \end{cases}$$

  - P(X < 6)?

$$P(x < 6) = \int_0^6 \frac{1}{20}\,dx = \frac{6}{20}$$

  - P(4 < X < 17)?

$$P(4 < x < 17) = \int_4^{17} \frac{1}{20}\,dx = \frac{17}{20} - \frac{4}{20} = \frac{13}{20}$$

Riding the Marguerite

Riding the Marguerite

- Say the Marguerite bus stops at the Gates bldg. at 15 minute intervals (2:00, 2:15, 2:30, etc.)
  - Passenger arrives at stop uniformly between 2-2:30pm
  - X ~ Uni(0, 30)
- P(Passenger waits < 5 minutes for bus)?
  - Must arrive between 2:10-2:15pm or 2:25-2:30pm

$$P(10 < X < 15) + P(25 < x < 30) = \int_{10}^{15} \frac{1}{30}\,dx + \int_{25}^{30} \frac{1}{30}\,dx = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$

- P(Passenger waits > 14 minutes for bus)?
  - Must arrive between 2:00-2:01pm or 2:15-2:16pm

$$P(0 < X < 1) + P(15 < x < 16) = \int_{0}^{1} \frac{1}{30}\,dx + \int_{15}^{16} \frac{1}{30}\,dx = \frac{1}{30} + \frac{1}{30} = \frac{1}{15}$$

When to Leave for Class

- Biking to a class on campus
  - Leave *t* minutes before class starts
  - X = travel time (minutes). X has PDF: $f(x)$
  - If early, incur cost: c/min. If late, incur cost: k/min.

$$\text{Cost}: \ C(X,t) = \begin{cases} c(t-X) & \text{if } x < t \\ k(X-t) & \text{if } x \geq t \end{cases}$$

  - Choose *t* (when to leave) to minimize E[C(*X*, *t*)]:

$$E[C(X,t)] = \int_0^\infty C(X,t)\, f(x)dx = \int_0^t c(t-x)\, f(x)dx + \int_t^\infty k(x-t)\, f(x)dx$$

Minimization via Differentiation

- Want to minimize w.r.t. *t*:

$$E[C(X,t)] = \int_0^t c(t-x)\, f(x)dx + \int_t^\infty k(x-t)\, f(x)dx$$

  - Differentiate E[C(*X*, *t*)] w.r.t. *t*, and set = 0 (to obtain *t**):
    - Leibniz integral rule:

$$\frac{d}{dt}\int_{f_1(t)}^{f_2(t)} g(x,t)dx = \frac{df_2(t)}{dt}\, g(f_2(t),t) - \frac{df_1(t)}{dt}\, g(f_1(t),t) + \int_{f_1(t)}^{f_2(t)} \frac{\partial g(x,t)}{\partial t} dx$$

$$\frac{d}{dt} E[C(X,t)] = c(t-t)\, f(t) + \int_0^t c f(x)dx - k(t-t)\, f(t) - \int_t^\infty k f(x)dx$$

$$0 = cF(t^*) - k[1 - F(t^*)] \;\Rightarrow\; F(t^*) = \frac{k}{c+k}$$