EGM101 – Skills Toolbox

Week 7, Part 5: Continuous Probability Distributions

Probability as Area

- Note that each "bar" of this chart has an area, A_i:

$$A_i = w_i H_i \longrightarrow A_i = w_i P(X = x_i)$$

- Because $w_i = 1$:

$$A_i = P(X = x_i)$$

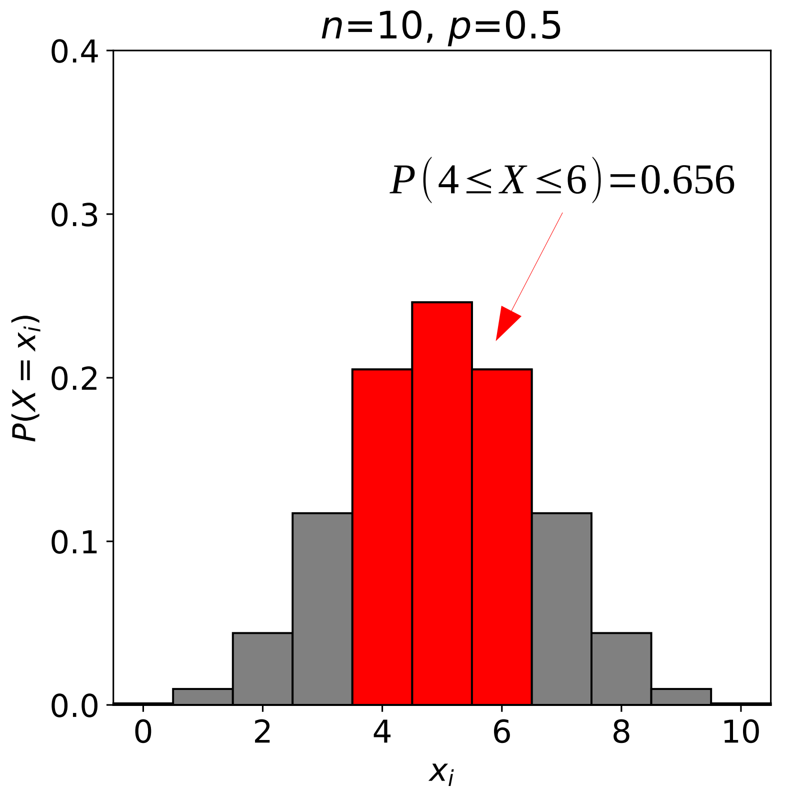
- The *total* area is:

$$A_t = \sum_{i=1}^{n} P(X = x_i) = 1$$

- And the area for some range is:

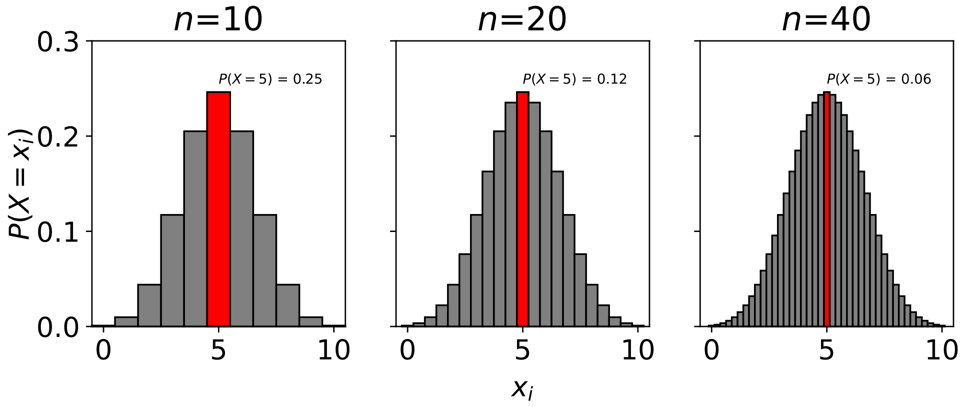
$$A_{j,k} = \sum_{i=j}^{k} P(X = x_i)$$

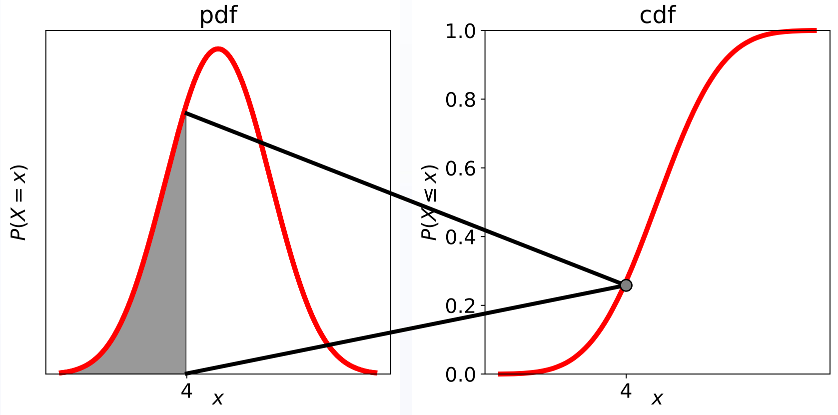
- What happens when we have continuous variables?



Continuous Probability Distributions

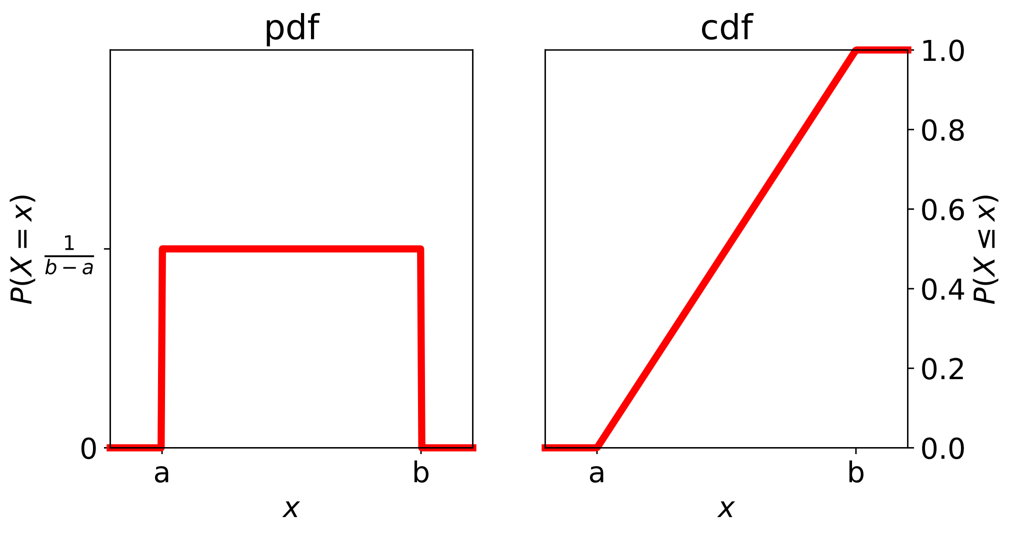
- Recall: a continuous variable can take on any value
  - i.e., the width of the bars gets very, very small
  - Probability of any specific value also gets very, very small
- So, probability is represented by area under a curve:
  - Probability density function (pdf), $f(x)$
  - To estimate area, use the cumulative distribution function (cdf)
  - Calculate probability for an *interval* of values
- Note that the cdf also tells us what proportion of values are below (or above) a given value





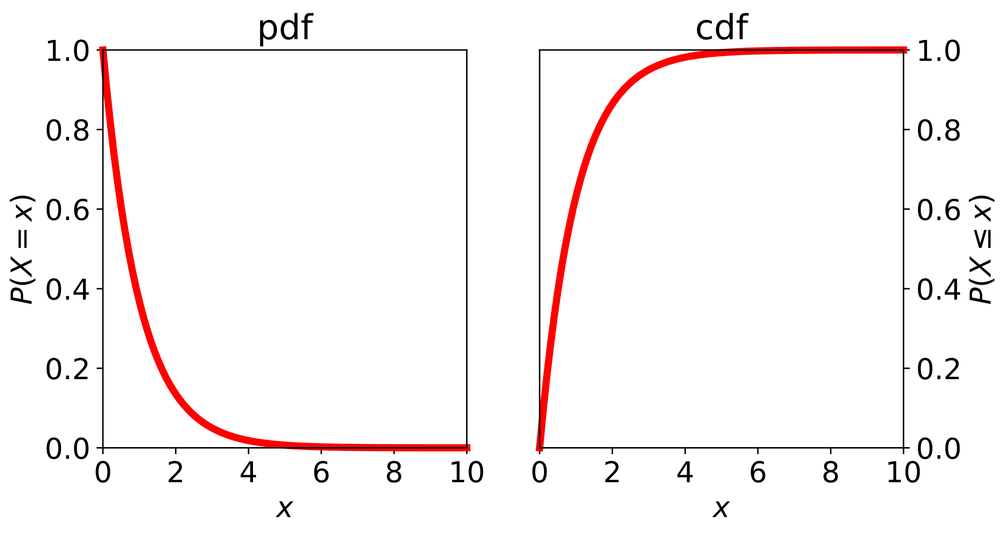
The (Continuous) Uniform Distribution

- All values within some range a, b have equal probability
- Important to note whether a, b are inclusive!
- Outside of a, b:
  - pdf(x) = 0
  - cdf(x) = 0 ($x < a$), 1 ($x > b$)



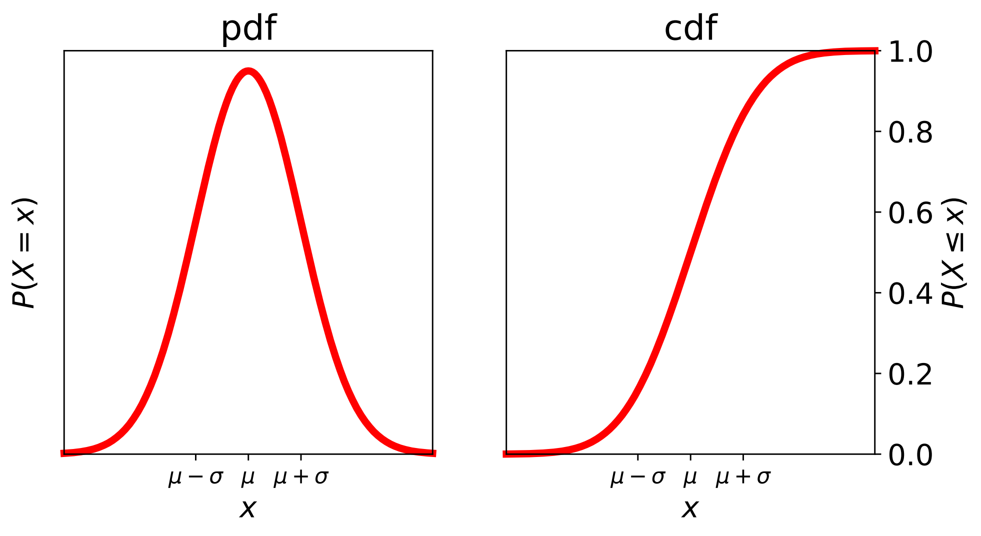
The Exponential Distribution

- Question: how long until the next event happens?
  - e.g., if a battery lasts 12 months on average, how likely is a new battery to last more than 8 months?
- Exponential distribution estimates probability using:
  - Mean "waiting time", μ
  - Decay parameter, $m = 1/\mu$
- Assumption: events occur continuously, independently at a constant rate
- Memoryless: "waiting time" does not depend on how much time has passed



The Normal Distribution

- Still the most important distribution for statistics/probability*
- Need to know:
  - Mean value, μ
  - Standard deviation, σ
- Properties:
  - Symmetrical (mean = mode = median)
  - As we move away from the mean (either direction), probability decreases quickly



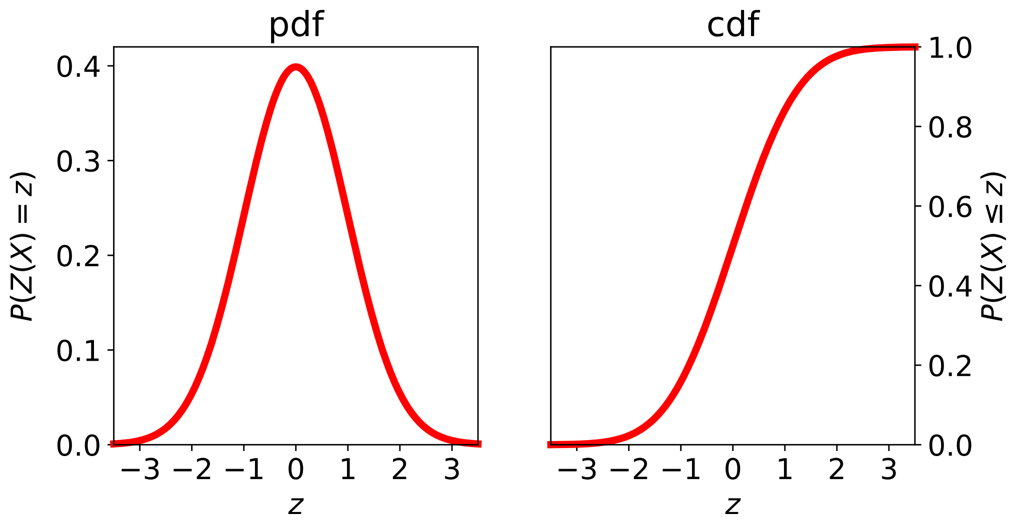
*probably.

Z-scores

- How best to compare for different variables?
- Define:

$$z = \frac{x - \mu}{\sigma}$$

- Tells us how many standard deviations x is from the mean
- Standard normal distribution:
  - $\mu = 0$, $\sigma = 1$
  - Estimate probability in terms of z
  - Enables comparison of differently scaled data
- In ancient times, people used statistical tables to look up cdf values based on z
  - Now, we mostly use computers for this.



The Empirical Rule

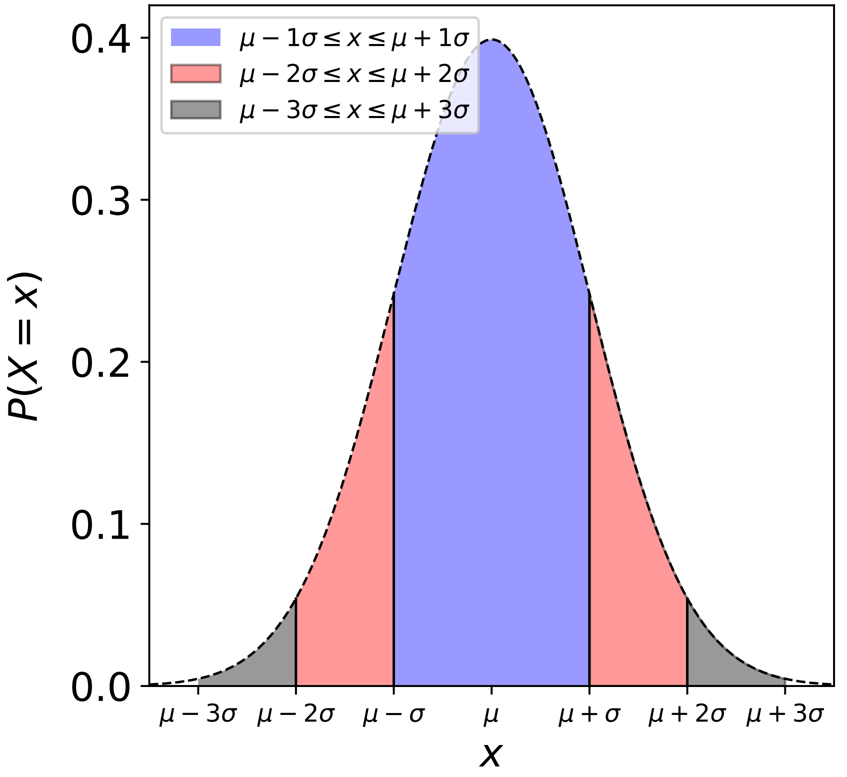
- If our data are normally distributed:

$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\%$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$$

- In other words:
  - 68.27% of x values within 1σ of mean
  - 95.45% of x values within 2σ of mean
  - 99.73% of x values within 3σ of mean
- Also known as the "68-95-99.7" rule



Example: The Fish Question

- Question (W7, P1, Slide 1): what is the likelihood of catching a fish larger than 100 cm?
  - Mean: 70 cm
  - Standard deviation: 15 cm
- Calculate the z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{100 - 70}{15} = 2$$

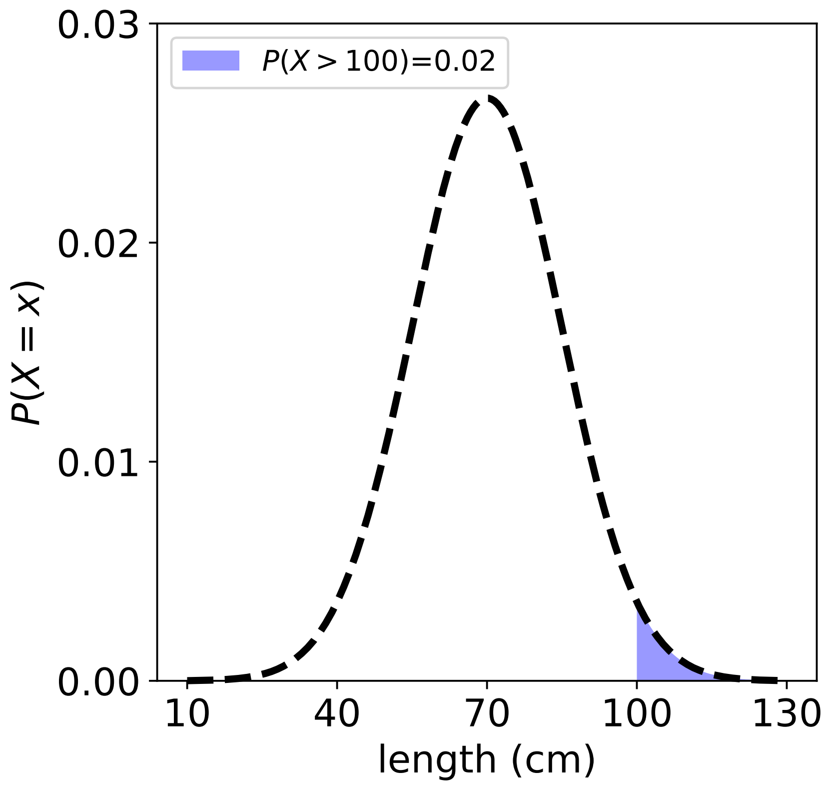
- Remember:

$$P(X \leq x) + P(X > x) = 1 \rightarrow P(X > x) = 1 - P(X \leq x)$$

- So:

$$P(X > 100) = 1 - cdf(100)$$

$$P(Z > 2) = 1 - cdf(2) \approx 0.0227$$



Summary

- With continuous variables, we have to think about ranges instead of individual values
- A number of common continuous distributions:
  - Uniform: for equal likelihood scenarios
  - Exponential: estimating time between events
  - Normal: for everything*
- With normal distribution, often consider z-score (# of standard deviations away from mean value)

*again, we assume this at our own peril

Additional Resources

- Illowsky and Dean, Chapters 5, 6
- Caswell, Chapters 8.9, 13.5–13.9
- Weiss, Chapter 6
- Uniform Distribution [Jim Frost]
- Exponential Distribution [Jim Frost]
- Empirical Rule [Jim Frost]
- Continuous probability distribution [Khan Academy]