Review of Probability and Statistics

SYS 611: Systems Modeling and Simulation

Paul T. Grogan, Ph.D.
Assistant Professor
School of Systems and Enterprises

Agenda

1. Samples and Statistics
2. Discrete Random Variables
3. Continuous Random Variables
4. Confidence Intervals

Reference: S.M. Ross, "Elements of Probability," "Random Numbers," Ch. 2-3 in *Simulation*, 2012.

J.V. Farr, "Review of Probability and Statistics," Ch. 3 in Simulation of Complex Systems and Enterprises, Stevens Institute of Technology, 2007.

Samples and Statistics

Samples and Statistics

A **sample** is an observation from a population

- Usually, cannot observe an entire population
- Each sample is subject to variation:
  - Aleatory variability: inherent uncertainty in process (randomness) – characterized by *random variables* (RVs)
  - Measurement error: calibration, environmental noise, instrument malfunction, human error, etc.

Statistics infer properties about a population

Descriptive Statistics

Descriptive statistics summarize observed samples and estimate properties about the population

- Sample mean $\bar{x}$ (estimate of population mean μ_x)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample variance s^2 (estimate of population variance σ_x^2)

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Sample standard deviation $s_x = \sqrt{s_x^2}$

Probability Basics

Probability quantifies chance of observing an event

- Observation of an event is either true (1) or false (0)
- What is the probability of a dice roll?
  - Outcome events: $D_1, D_2, D_3, D_4, D_5, D_6, D_{odd}, D_{even}$
  - Event probability: $P(D_6) = ?$ $P(D_{odd}) = ?$
- What is the probability of tomorrow's weather?
  - Outcome events: $W_{clear}, W_{cloudy}, W_{rain}, W_{snow}$
  - Event probability: $P(W_{rain}) = ?$ $P(W_{snow}) = ?$

Discrete Random Variables

Random Variables

Random variables assign events to numbers

- **Discrete random variables** assign events to *countable* numbers (e.g., integers)
  - Elementary events
  - Mutually exclusive and collectively exhaustive
- What is the *probability* of a dice roll?
  - Outcome event: X: random variable (value of roll)

$P\{X = 1\} = 1/6$ $P\{X = 2\} = 1/6$ $P\{X = 3\} = 1/6$

$P\{X = 4\} = 1/6$ $P\{X = 5\} = 1/6$ $P\{X = 6\} = 1/6$

$P\{X = 0\} = 0$ $P\{X = 7\} = 0$ $P\{X = -1\} = 0$

Example: Discrete R.V.

How many heads from flipping a fair coin three times?

$$X = \text{number of heads}$$

Sample space:

Event	$P(E)$	X
TTT	1/8	0
TTH	1/8	1
THT	1/8	1
THH	1/8	2

Event	$P(E)$	X
HTT	1/8	1
HTH	1/8	2
HHT	1/8	2
HHH	1/8	3

Distribution:

x	$P\{X = x\}$
0	1/8
1	3/8
2	3/8
3	1/8

Probability Mass Functions

Probability Mass Function (PMF) maps *discrete* random variables to probability *masses*

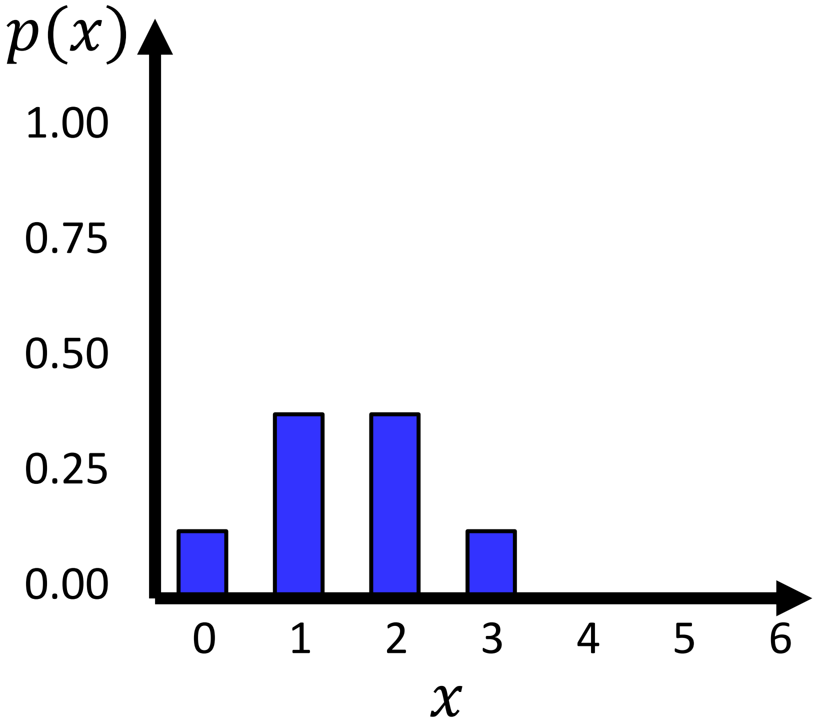
- Functional notation:
  - $P\{X = x\} = p(x)$
  - X: random variable (number of heads in 3 flips)
  - x: event outcome variable
- Note: $\sum_{x=-\infty}^{\infty} p(x) = 1$

$\boldsymbol{x}$	$\boldsymbol{p(x)}$
-1	0
0	1/8
1	3/8
2	3/8
3	1/8
4	0

PMF Plots

- PMF plots are *analytical* histograms
  - Replace frequency with probability mass

x	$p(x)$
-1	0
0	1/8
1	3/8
2	3/8
3	1/8
4	0



Cumulative Distribution Funct.

Cumulative Distribution Function (CDF) maps random variable ranges to probabilities

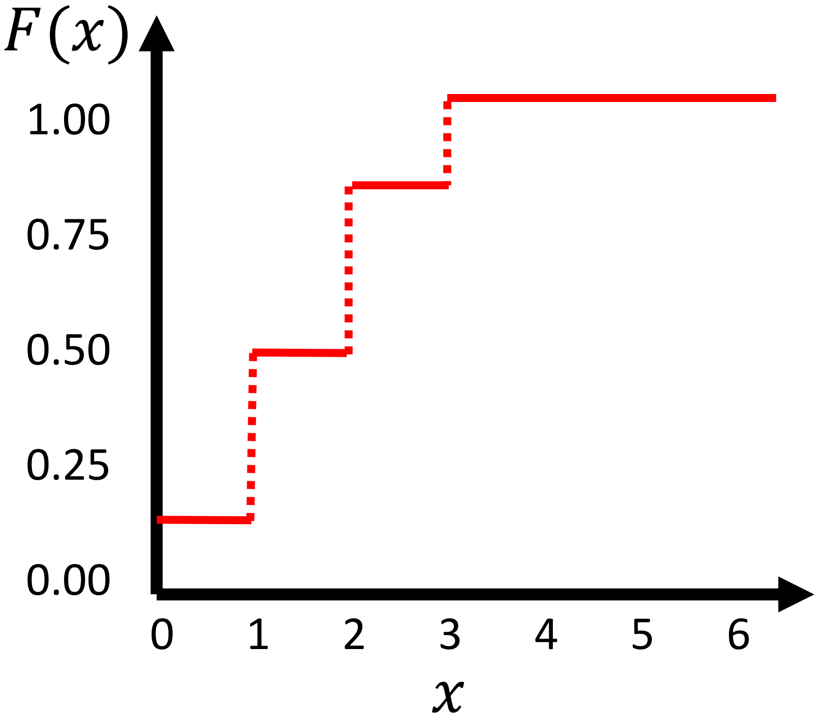
- Functional notation:
  - $P\{X \leq x\} = F(x)$
  - X: random variable (number of heads in 3 flips)
  - x: event outcome variable
- Note: $F(x) = \sum_{i=0}^{x} p(i)$

x	$p(x)$	$F(x)$
-1	0	0
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	8/8
4	0	8/8

CDF Plot

- CDF plots are *analytical* cumulative frequency plots
  - Replace cumulative freq. with cumulative prob.

x	$p(x)$	$F(x)$
-1	0	0
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	8/8
4	0	8/8



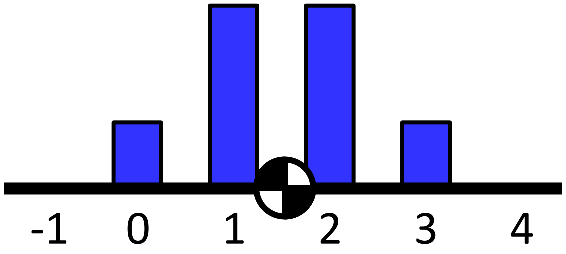
Mean or Expected Value

Expected value of a discrete distribution:

$$\mu_x = E[X] = \sum_{x=-\infty}^{\infty} x \cdot p(x)$$

Analogous to first moment (center of mass)

$$\sum_{x=0}^{3} x \cdot p(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2}$$



Variance or Standard Deviation

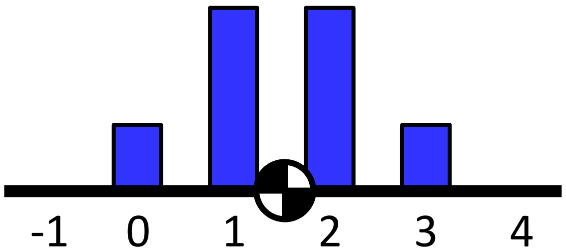
Variance of a discrete distribution:

$$\sigma_x^2 = Var(X) = E[(X - E[X])^2] = \sum_{x=-\infty}^{\infty} (x-\mu)^2 \cdot p(x)$$

$$= E[X^2] - (E[X])^2 = \sum_{x=-\infty}^{\infty} x^2 \cdot p(x) - \mu_x^2$$

Analogous to discrete form of moment of inertia

$$\sum_{x=0}^{3} (x-\mu)^2 \cdot p(x) = \left(0 - \frac{3}{2}\right)^2 \cdot \frac{1}{8} + \left(1 - \frac{3}{2}\right)^2 \cdot \frac{3}{8} + \left(2 - \frac{3}{2}\right)^2 \cdot \frac{3}{8} + \left(3 - \frac{3}{2}\right)^2 \cdot \frac{1}{8} = \frac{9 \cdot 1 + 1 \cdot 3 + 1 \cdot 3 + 9 \cdot 1}{4 \cdot 8} = \frac{3}{4}$$



Common Discrete Distributions

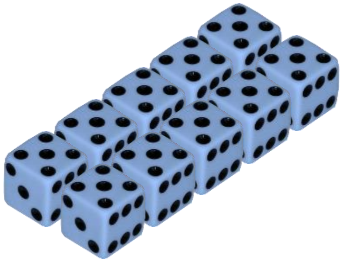
Name	Parameters	$p(x)$	$\mu_x = E[X]$	$\sigma_x^2 = Var(X)$
Uniform	Min $a \geq 0$, Max $b > a$, Num. $n > 0$	$\frac{1}{n}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Binomial	Trials $n > 0$, Prob. $p > 0$	$\binom{n}{x} p^x (1-p)^{n-x}$	$n \cdot p$	$n \cdot p \cdot (1-p)$
Geometric	Prob. $p > 0$	$p(1-p)^{x-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	Rate $\lambda > 0$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ
Hyper-geometric	Trials $n > 0$, Pop. $N > 0$, Succ. $0 < A < N$	$\frac{\binom{A}{x} \cdot \binom{N-A}{n-x}}{\binom{N}{n}}$	$n \cdot \frac{A}{N}$	$n \cdot \frac{A}{N} \cdot \frac{N-A}{N} \cdot \frac{N-n}{N-1}$
Negative Binomial	Successes $n > 0$, Prob. $p > 0$	$\binom{x+n-1}{n-1} p^n (1-p)^x$	$\frac{n \cdot (1-p)}{p}$	$\frac{n \cdot (1-p)}{p^2}$

Dice Roller Experiment

- The team has 10 dice and hits on a {3,4,5,6}.
- What is the probability of exactly $Y = y$ hits?
  - Y: # hits in 10 dice
  - $p(y) = P\{Y = y\}$
- Virtual dice: **random.org/dice**
- Post results: **pollev.com/pgrogan**

Blue Team:

- Small fighting force
- 4x effective weapons



- Roll 3|4|5|6 to hit target

Exercise: Dice Roller

- The team has 10 dice and hits on {3, 4, 5, 6}.
- What is the PMF/CDF for the number of hits?

$$Y\text{: number of hits from 10 dice,} \qquad P\{Y = y\} = p(y)$$

$$X_i\text{: } i\text{th die scores a hit,} \qquad P(X_i) = \frac{4}{6} = \frac{2}{3}$$

$$p(10) = \binom{10}{10} \cdot P(X_1) \cdot \ldots \cdot P(X_{10}) = \left(\frac{4}{6}\right)^{10}$$

$$\Rightarrow p(y) = \binom{10}{y}\left(\frac{2}{3}\right)^{y}\left(1 - \frac{2}{3}\right)^{10-y}$$

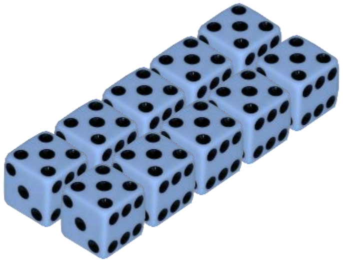
$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad Y \sim \text{binomial}\left(n = 10, p = \frac{2}{3}\right)$$

Critical Hit Experiment

- The team has 10 dice and hits on a {3,4,5,6}.
- Y: # hits in 10 dice
- $Y \geq 9$ is a "critical hit"
- What is the probability of observing $K = k$ "critical hits" in 10 seconds?
  - K: # critical hits in 10 sec
  - $p(k) = P\{K = k\}$

Blue Team:

- Small fighting force
- 4x effective weapons



- Roll 3|4|5|6 to hit target

Exercise: Critical Hit

- Number of "critical hits" in 10 seconds is a Poisson process:

$$K \sim \text{poisson}(\lambda), \qquad p(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

- Probability of "critical hit" is:

$$P(\text{Crit}) = \left(\frac{4}{6}\right)^{10} + 10\left(\frac{4}{6}\right)^{9}\left(\frac{2}{6}\right) = \frac{1024 + 5120}{59049} = \frac{2048}{19683} = 0.104$$

- N students rolling dice, each roll takes T seconds
- Expect long-term average "critical hit" rate to be

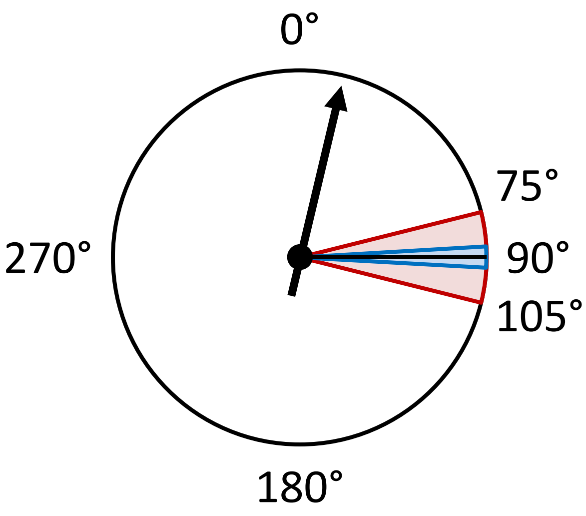
$$\lambda = P(\text{Crit}) \cdot \frac{N}{T} \text{ per second} = \frac{P(\text{Crit}) \cdot N}{(T/10)} \text{ per 10 seconds}$$

Continuous Random Variables

Continuous Random Variables

Random variables assign events to numbers

- Continuous random variables assign events to *uncountable* numbers (e.g., floating-point)
- What is the probability of a spinner stopping at a certain angle?
  - Outcome event: X: random variable (angle in degrees)
  - Event probability:
    - $P\{75.000 \leq X \leq 105.000\} = \frac{30}{360} = \frac{1}{12}$
    - $P\{85.000 \leq X \leq 95.000\} = \frac{10}{360} = \frac{1}{36}$
    - $P\{X = 90.000\} = 0$



Probability Density Functions

Probability Density Function (PDF) maps *continuous* random variables to probability *densities*

- Functional notation:
  - $f(x) \approx P\{x \le X \le x + \Delta x\}/\Delta x$
  - X: random variable
  - x: event outcome
  - Δx: small range of outcome (volume)
- Magnitude of $f(x)$ is selected such that

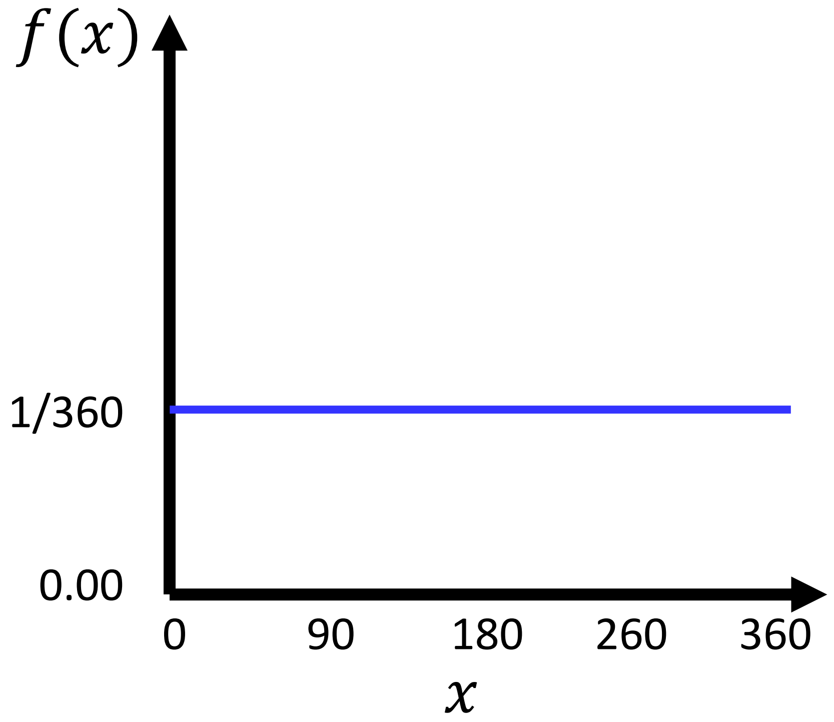
$$\int_{-\infty}^{\infty} f(x)dx = 1, \qquad \int_{0}^{360} f(x)dx = 1 \Rightarrow f(x) = 1/360$$

x	$f(x)$
0.0	1/360
10.0	1/360
23.4	1/360
89.9	1/360
90.0	1/360
180.0	1/360
359.9	1/360
360.1	0

PDF Plots

- PDF plots are like histograms, but use line charts instead of bar charts

x	$f(x)$
0.0	1/360
10.0	1/360
23.4	1/360
89.9	1/360
90.0	1/360
180.0	1/360
359.9	1/360
360.1	0



Cumulative Distribution Funct.

Cumulative Distribution Function (CDF) maps random variable ranges to probabilities

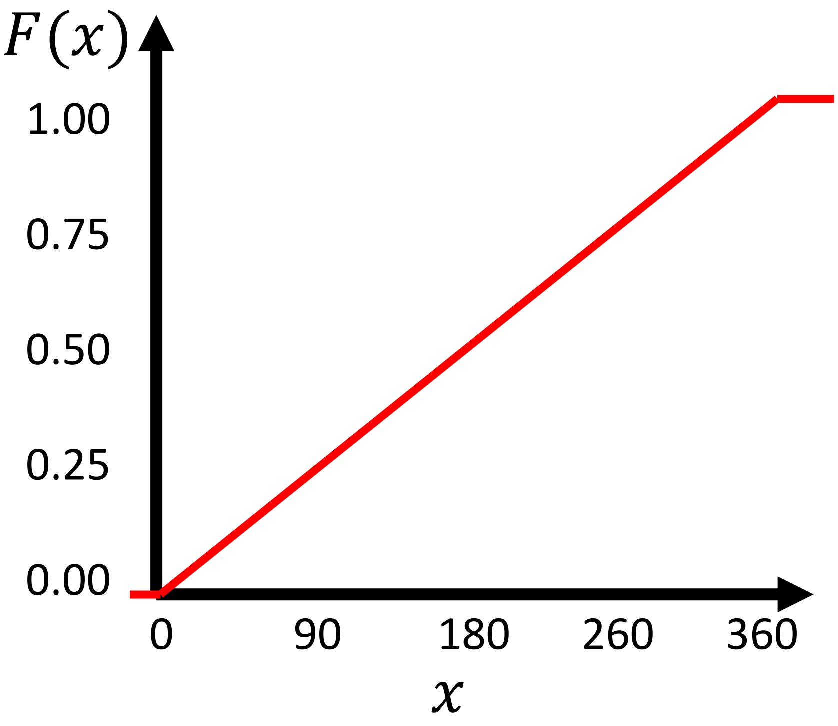
- Functional notation:
  - $P\{X \leq x\} = F(x)$
  - X: random variable
  - x: event outcome
- $F(x) = \int_{-\infty}^{x} f(i) di$

x	$f(x)$	$F(x)$
0.0	1/360	0.0/360
10.0	1/360	10.0/360
23.4	1/360	23.4/360
89.9	1/360	89.9/360
90.0	1/360	90.0/360
180.0	1/360	180/360
359.9	1/360	359.9/360
360.1	0	1

CDF Plot

- CDF plots are like cumulative frequency plots
  - Replace cumulative freq. with cumulative prob.

x	$f(x)$	$F(x)$
0.0	1/360	0.0/360
10.0	1/360	10.0/360
23.4	1/360	23.4/360
89.9	1/360	89.9/360
90.0	1/360	90.0/360
180.0	1/360	180/360
359.9	1/360	359.9/360
360.1	0	1



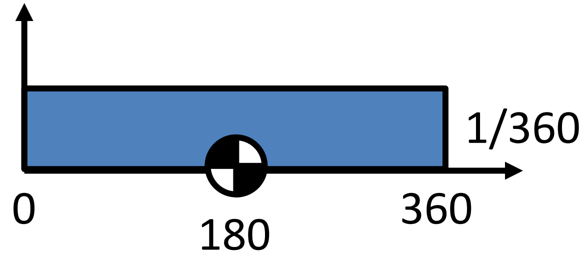
Mean or Expected Value

Expected value of a continuous distribution:

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

Analogous to the first moment (center of mass)

$$\int_{-\infty}^{\infty} x \cdot f(x) \cdot dx = \int_{0}^{360} x \cdot \frac{1}{360} \cdot dx = \frac{1}{360} \cdot \frac{x^2}{2}\bigg|_0^{360} = \frac{360^2 - 0^2}{360 \cdot 2} = 180$$



Variance or Standard Deviation

Variance of a continuous distribution:

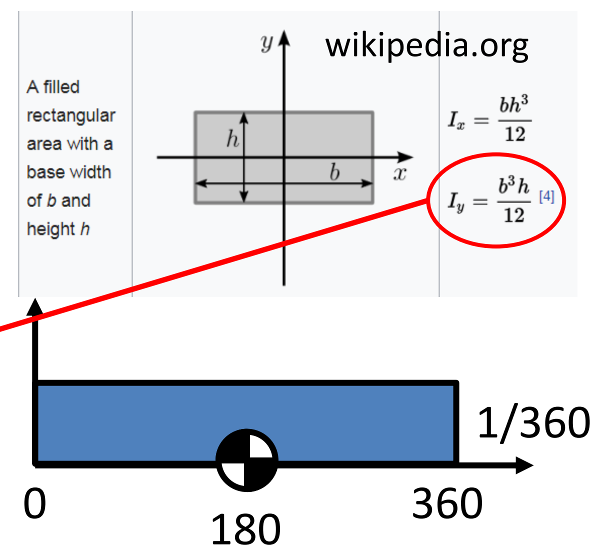
$$\sigma_x^2 = Var(x) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x) \cdot dx$$

$$= E[X^2] - E[X]^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x) \cdot dx - \mu_x^2$$

Analogous to moment of inertia:

$$\int_0^{360} x^2 \cdot \frac{1}{360} \cdot dx - 180^2 = \frac{1}{360} \cdot \frac{x^3}{3}\Big|_0^{360} - 180^2$$

$$= \frac{360^2}{3} - 180^2 = 10800 = \frac{1}{12} \cdot 360^3 \cdot \frac{1}{360}$$



Common Continuous Distributions

Name	Parameters	$f(x)$	$\mu_x = E[X]$	$\sigma_x^2 = Var(X)$
Uniform	Min a, Max $b > a$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Ramp Up	Min a, Max $b > a$	$\frac{2(x-a)}{(b-a)^2}$	$\frac{a+2b}{3}$	$\frac{a^2-2ab+b^2}{18}$
Triangular	Min a, Mode $c > a$, Max $b > c$	$\begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & a \leq x < c \\ \frac{2(b-x)}{(b-a)(b-c)} & c < x < b \end{cases}$	$\frac{a+b+c}{3}$	$\frac{a^2+b^2+c^2-ab-ac-bc}{18}$
Normal	Mean μ, Std. Dev. σ	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Standard Normal	(none)	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	0	1

Common Continuous Distributions

Name	Parameters	$f(x)$	$\mu_x = E[X]$	$\sigma_x^2 = Var(X)$
Chi-squared	Degrees of freedom k	$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	k	$2k$
Exponential	Rate λ	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Erlang	Rate λ, Shape k	$\frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$
Gamma	Rate λ, Shape k	$\frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$
Weibull	Rate λ, Shape α	$\alpha \lambda^\alpha x^{\alpha-1} e^{-(\lambda x)^\alpha}$	$\frac{1}{\lambda}\Gamma\left(1+\frac{1}{\alpha}\right)$	$\frac{1}{\lambda^2}\left[\Gamma\left(1+\frac{2}{\alpha}\right)+\left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2\right]$

Exercise: Critical Hit

- Previously studied the number of events per time:
  - K: number of "critical hits" in 10 seconds
  - λ: average rate of "critical hit" events (per 10 seconds)

$$K \sim \text{poisson}(\lambda), \qquad p(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

- Now study the time between adjacent events:
  - T: time between "critical hit" events
  - λ: average rate (per 10 seconds) of "critical hit" events

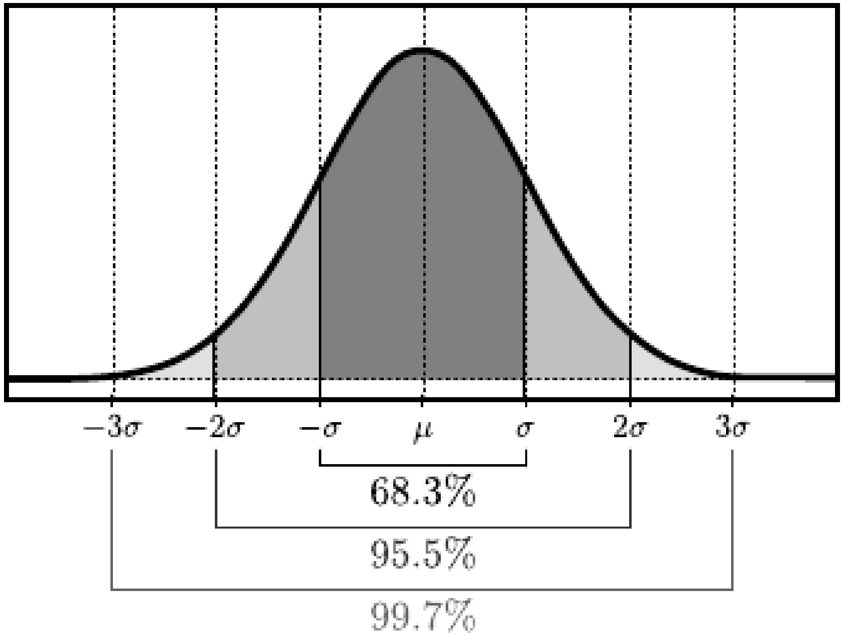
$$T \sim \text{exponential}(\lambda), \qquad f(t) = \lambda \cdot e^{-\lambda \cdot t}, \qquad F(t) = 1 - e^{-\lambda \cdot \mathrm{t}}$$

Confidence Intervals

Normal Distribution

Normal distribution (Gaussian) describes some natural random variables (e.g., human height)

- Mathematically describes the summation of many random variables
  - Used in SYS 611 to model distribution of sample means
  - Collect 1 sample mean but infer population statistics



Central Limit Theorem

Central Limit Theorem (CLT) states the sample mean of independent samples approaches a normal distribution regardless of the source population distribution.

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables, each with finite mean μ_x and variance σ_x^2. Then the distribution of

$$\frac{(X_1 + X_2 + \cdots + X_n)}{n} = \bar{X}$$

approaches a **normal distribution** as $n \to \infty$.

$$\bar{X} \sim \text{normal}\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

Dice Roller Sample Mean

$$X \sim \text{binomial}(n, p)$$

$$\mu_x = n \cdot p$$

$$\sigma_x = \sqrt{n \cdot p \cdot (1 - p)}$$

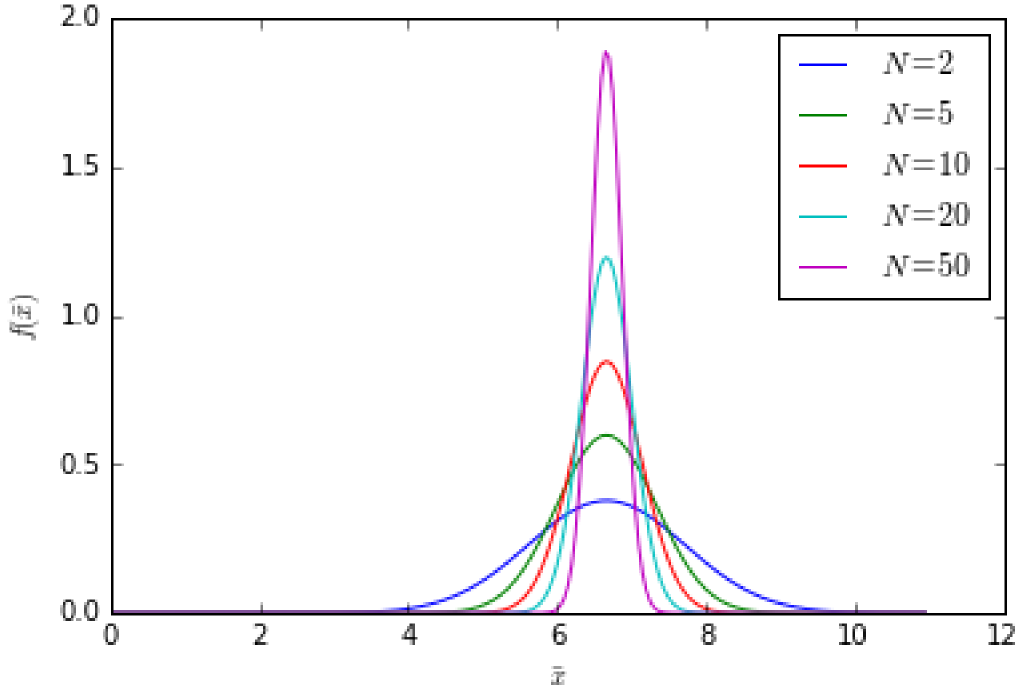
Sample mean from N samples:

$$\bar{X} \sim \text{normal}(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

$$\mu_{\bar{x}} = \mu_x = n \cdot p = 10 \cdot \frac{2}{3} = 6.67$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} = \frac{\sqrt{(n \cdot p \cdot (1 - p)}}{\sqrt{N}}$$

$$= \frac{\sqrt{10 \cdot 2/3 \cdot 1/3}}{\sqrt{N}} = \frac{1.49}{\sqrt{N}}$$



Confidence Intervals

Confidence intervals apply the CLT to infer the population mean based on a number of samples:

- $(1 - \alpha) * 100\%$ confidence interval:

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{N}}$$

- Critical z-score: $z_{1-\alpha/2}$

$$z_{0.975} = \text{normal}^{-1}(0.975,0,1) = 1.96$$

- Standard error of mean (SEM): $\frac{\sigma_x}{\sqrt{N}} \approx \frac{s_x}{\sqrt{N}}$

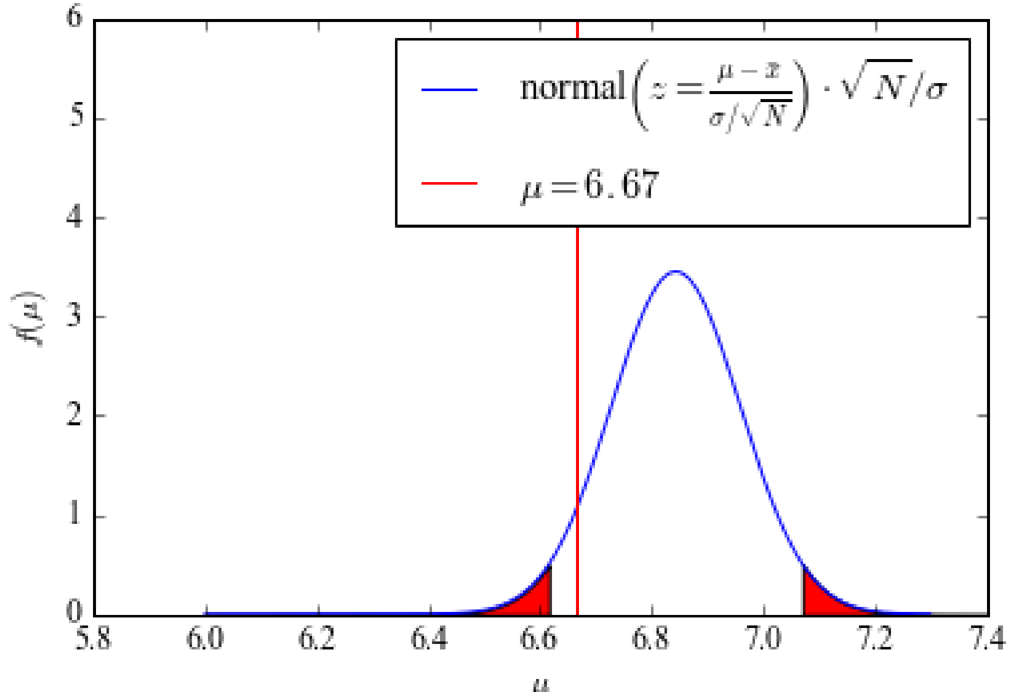
Estimating Population Mean

- Consider 166 samples with following statistics:
- $\bar{x} = 6.84,\ s_x = 1.49$
- $\mu_x \in \bar{x} \pm z_{1-\alpha/2} \frac{s_x}{\sqrt{N}}$

$$= 6.84 \pm 1.96 \cdot \frac{1.49}{\sqrt{166}}$$

$$= [6.62, 7.07]$$

- Population mean will fall in the range $[6.62, 7.07]$ 95% of the time



$$\bar{x} - z_{0.975} \frac{s_x}{\sqrt{N}} = 6.62$$

$$\bar{x} + z_{0.975} \frac{s_x}{\sqrt{N}} = 7.07$$