Review of Probability and Statistics

SYS-611: Simulation and Modeling

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

Reading: 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

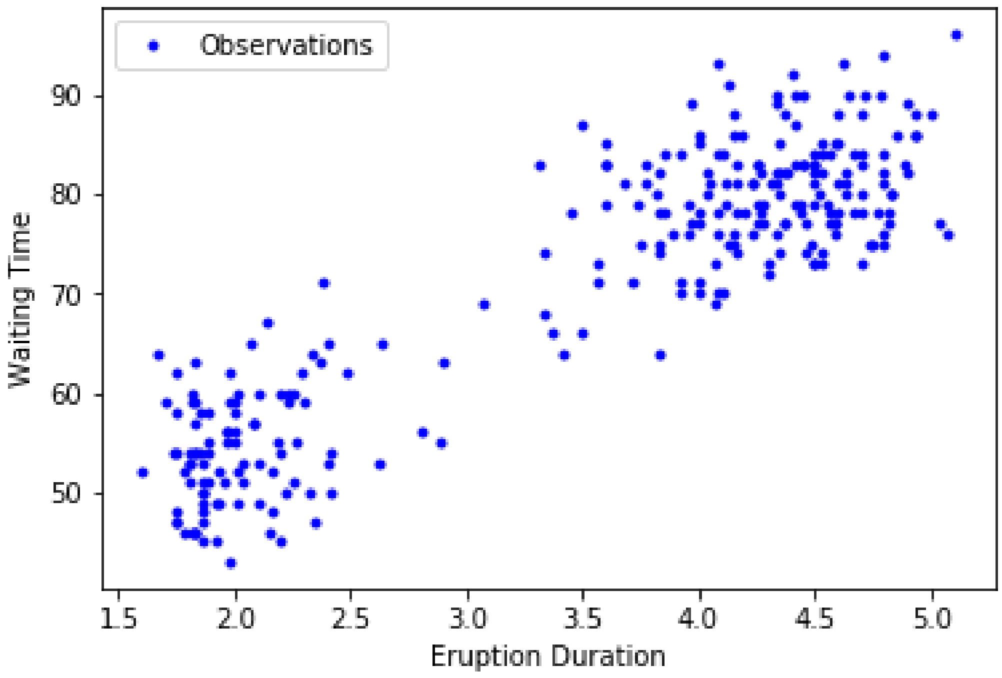
- A **sample** is an observation of an event
- Samples are subject to:
  - Aleatory variability: inherent uncertainty in process
  - Epistemic uncertainty: prediction error due to limitations in measurement or incorrect theoretical models
  - Measurement error: calibration bias, environmental noise, instrument malfunction, human error, etc.
- **Statistics** characterize a sample population

Old Faithful Dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

df = pd.read_csv('faithful.csv')

plt.figure()
plt.plot(
     df['eruptions'],
     df['waiting'],
     '.b',
     label='Observations'
)
plt.xlabel('Eruption Duration')
plt.ylabel('Waiting Time')
plt.legend(loc='best')
```



What are some sources of variability in this dataset?

Descriptive Statistics

Descriptive statistics summarize the population from which samples are observed

- Sample mean $\bar{x}$ (→ population mean μ)

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

- Sample variance s^2 (→ population variance σ^2)

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

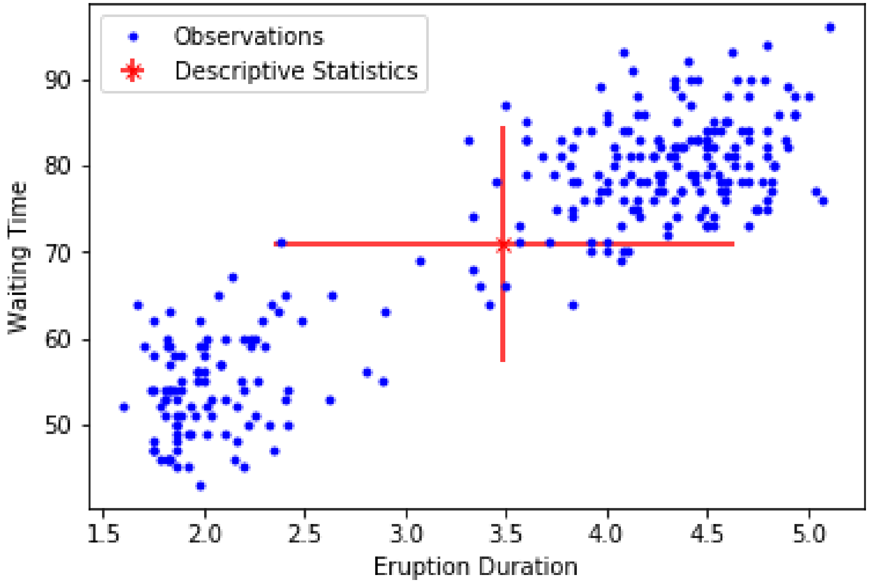
- Sample standard deviation $s = \sqrt{s^2}$

Old Faithful Statistics

```
dur_mean = np.mean(data[:,1])
dur_std = np.std(data[:,1], ddof=1)

wait_mean = np.mean(data[:,2])
wait_std = np.std(data[:,2], ddof=1)

plt.errorbar(dur_mean, wait_mean,
     fmt='xr',
     xerr=duration_std,
     yerr=wait_std,
     label='Descriptive Statistics'
)
plt.legend(loc='best')
```




```
>> dur_mean = 3.49
>> dur_std = 1.14
>> wait_mean = 70.90
>> wait_std = 13.57
```

How well do these statistics describe the Old Faithful data?

Regression Modeling

Regression models mathematical relationships for independent (x) and dependent (y) variables.

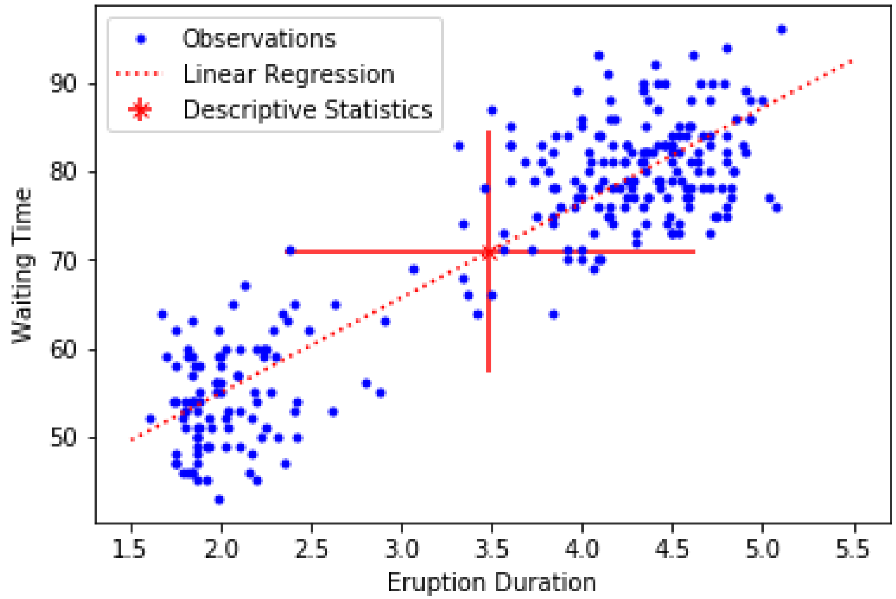
- General regression: $y = f(x)$
- Linear regression: $y = \beta_1 x + \beta_0$
- Polynomial regression: $y = \beta_k x^k + \cdots + \beta_1 x + \beta_0$
- Multiple regression: $y = \beta_n x_n + \cdots + \beta_1 x_1 + \beta_0$
- Coefficients are typically found using the least squares method (use a library function)
- Epistemic uncertainty is a new source of error between observations and predictions

Old Faithful Regression

$$y = 10.73x + 33.47 + \varepsilon$$

```
coefs = np.polyfit(
    df['eruptions'],
    df['waiting'],
    1
)
reg_x = np.array([1.5, 5.5])
reg_y = coefs[0]*reg_x + coefs[1]

plt.plot(reg_x, reg_y, ':r',
  label='Linear Regression'
)
plt.legend(loc='best')
```




```
>> coefs[0] = 10.73
>> coefs[1] = 33.47
```

How to interpret coefficients?

Aleatory variability (ε) vs. epistemic uncertainty (coefficients/model)?

Correlation Analysis

- **Correlation** measures the relationship between two variables without implying causality
- Correlation coefficient r
- Coefficient of determination r^2 : percent of variance explained by variable(s)

Example: Old Faithful Correlation

```
r, p = stats.pearsonr(df['eruptions'], df['waiting'])
>> r = 0.90
>> r**2 = 0.81
```

Discrete Random Variables

Probability Basics

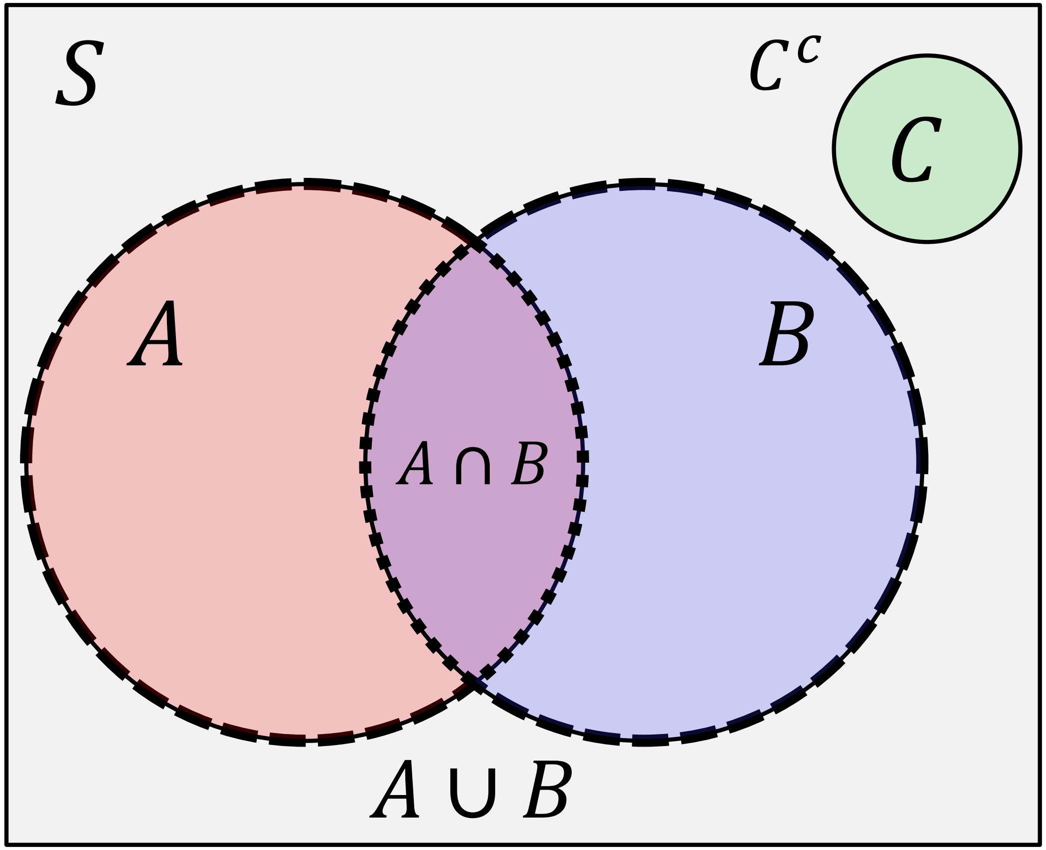
Probability quantifies the chance an event occurs

- Occurrence 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: $P(W_{rain}) = ?$ $P(W_{snow}) = ?$

Probability Operations

- **Union** $(A \cup B)$
- **Intersection** $(A \cap B)$
- **Complement** (C^c)
- **Law of Addition**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Law of Multiplication:** $P(A \cap B) = P(A) \cdot P(B|A)$ $= P(B) \cdot P(A|B)$
- **Conditional Prob.**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Visualization of probability space S:
probability of events A, B, C is proportional to area



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)

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

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

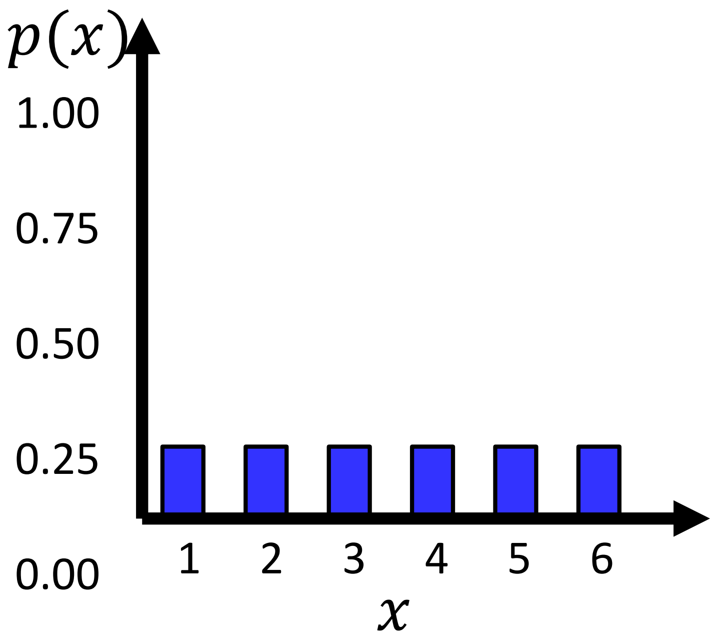
$$P\{X=0\}=0 \quad P\{X=7\}=0 \quad P\{X=-1\}=0$$

Probability Mass Functions

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

- Functional notation:
  - $P\{X = x\} = p(x)$
  - X: random variable (value of dice roll)
  - x: event outcome
- Note: $\sum_{x=0}^{\infty} p(x) = 1$

x	$p(x)$
0	0
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
7	0

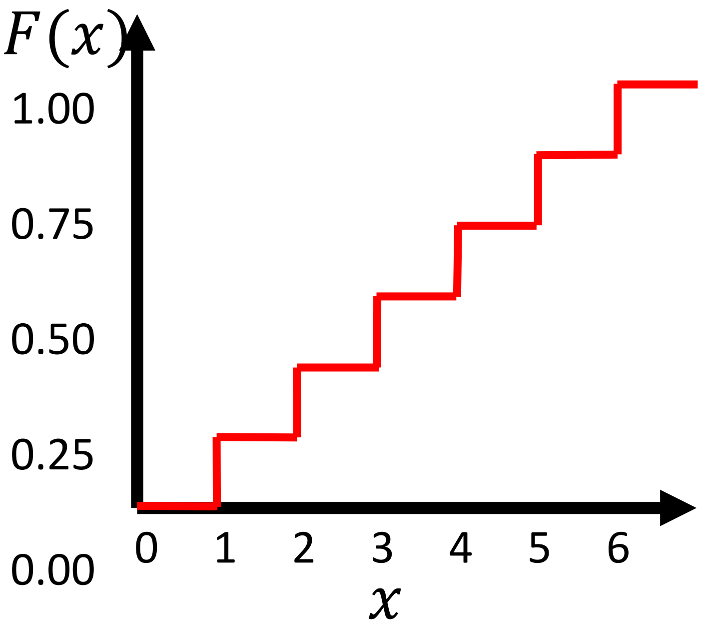


Cumulative Distribution Funct.

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

- Functional notation:
  - $P\{X \leq x\} = F(x)$
  - X: random variable (value of dice roll)
  - x: event outcome
- $F(x) = \sum_{i=0}^{x} p(i)$

x	$p(x)$	$F(x)$
0	0	0
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
7	0	1



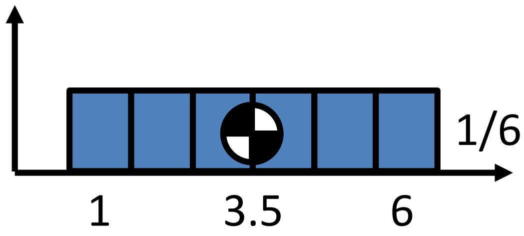
Mathematical Expectation

- **Expected value** of a discrete distribution:

$$\mu = E[X] = \sum_{x=0}^{\infty} x \cdot p(x)$$

- Analogous to first moment (center of mass)

$$\sum_{x=1}^{6} x \cdot p(x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

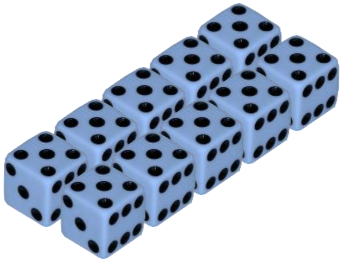


Dice Roller Activity

- 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
  - $p(y) = P\{Y = y\}$
- Let's collect some data: [goo.gl/orkF6d]
- Online students, use [random.org/dice]

Blue Team:

- Small fighting force
- 3x 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: number of hits from 10 dice, $P\{Y = y\} = p(y)$
  - X_i: dice i scores a hit, $P(X_i) = \frac{4}{6} = \frac{2}{3}$
  - $p(10) = \binom{10}{0} * P(X_1) * \cdots * P(X_{10}) = \left(\frac{4}{6}\right)^{10}$
  - $\rightarrow p(y) = \binom{10}{y}\left(\frac{2}{3}\right)^{y}\left(1 - \frac{2}{3}\right)^{10-y}$

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

Yahtzee Game

- Play a mini version of Yahtzee with one roll of 3 dice
- What is the probability of observing $Z = z$ mini-Yahtzees in 1 minute?
  - Z: # mini-Yahtzees in 1 min
  - $p(z) = P\{Z = z\}$
- Let's collect some data: goo.gl/ukyAU2
- Online students, use random.org/dice

Fair Use Image by Nanami

Exercise: Mini-Yahtzee

- Probability of “mini-Yahtzee” is:

$$P(\text{Yahtzee}) = \left(\frac{1}{1}\right) \cdot \left(\frac{1}{6}\right) \cdot \left(\frac{1}{6}\right) = \frac{1}{36}$$

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

$$\lambda = \frac{N}{36 \cdot T} \text{ per second} = \frac{N}{36 \cdot (T/60)} \text{ per minute}$$

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

Discrete Prob. Distributions

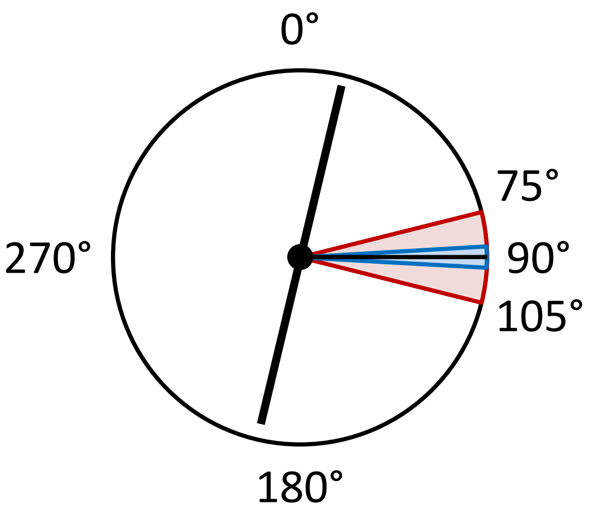
- **Binomial**: Models the number of successes in n independent trials with probability of success p.
- **Hypergeometric**: Models the number of successes in n independent trials without replacement from a population of size N with A possible successes
- **Negative Binomial**: Models the required number of independent trials to achieve n successes with probability of success p.
- **Poisson**: Models the number of independent events occurring in a fixed time or space with mean rate λ (occurrences/time or space).

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*

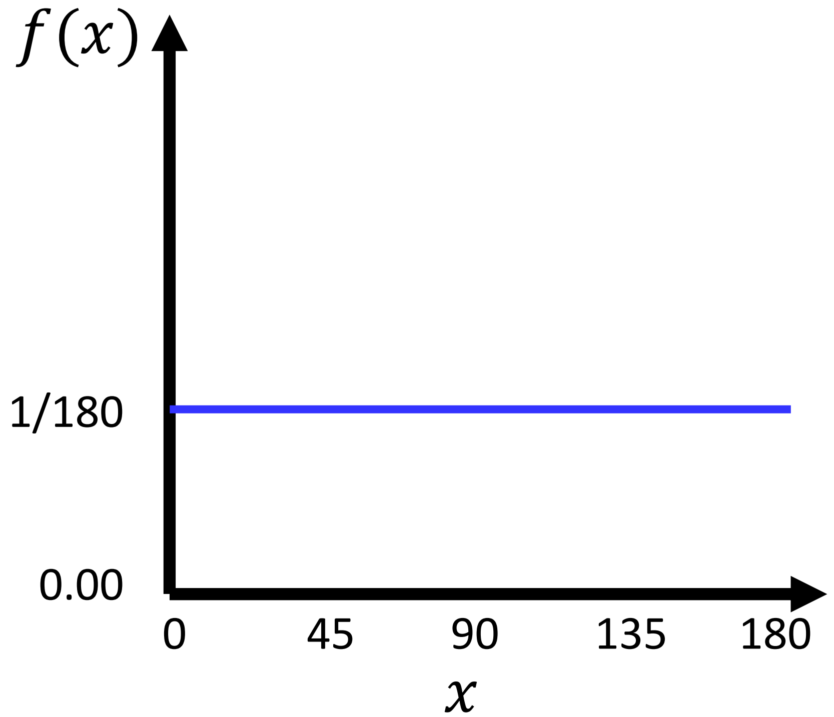
- Functional notation:
  - $P\{x \leq X \leq x + \Delta x\} \approx f(x) \cdot \Delta x$
  - X: random variable
  - x: event outcome
  - Δx: small range of outcome (volume)
- Note: $\int_{-\infty}^{\infty} f(x) dx = 1$

x	$f(x)$
0.0	1/180
10.0	1/180
23.4	1/180
44.9	1/180
45.0	1/180
90.0	1/180
179.9	1/180
181.0	0

PDF Plots

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

x	$f(x)$
0.0	1/180
10.0	1/180
23.4	1/180
44.9	1/180
45.0	1/180
90.0	1/180
179.9	1/180
181.0	0



Cumulative Distribution Funct.

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

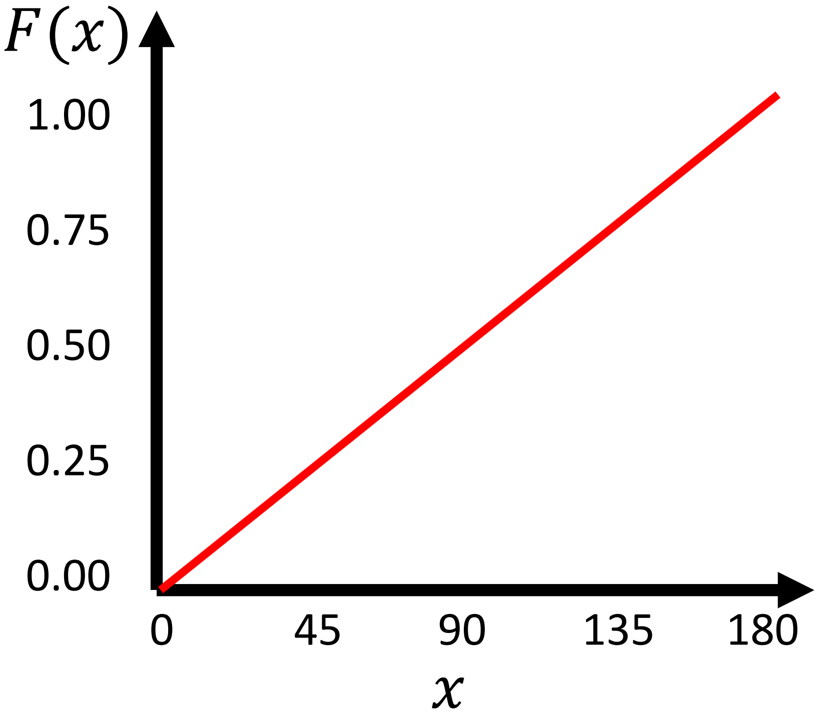
- Functional notation:
  - $P\{X \leq x\} = F(x)$
  - X: random variable
  - x: event outcome
- $F(x) = \int_{-\infty}^{x} f(i) di$

x	$f(x)$	$F(x)$
0.0	1/180	0.0/180
10.0	1/180	10.0/180
23.4	1/180	23.4/180
44.9	1/180	44.9/180
45.0	1/180	45.0/180
90.0	1/180	90.0/180
179.9	1/180	179.9/180
181.0	0	1

CDF Plot

- CDF plots similar to cumulative frequency plots
  - Replace cumulative freq. with cumulative prob.

x	$f(x)$	$F(x)$
0.0	1/180	0.0/180
10.0	1/180	10.0/180
23.4	1/180	23.4/180
44.9	1/180	44.9/180
45.0	1/180	45.0/180
90.0	1/180	90.0/180
179.9	1/180	179.9/180
181.0	0	1



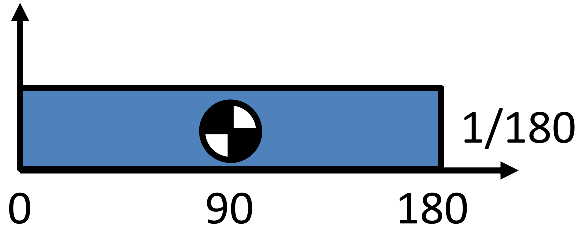
Mathematical Expectation

- **Expected value** of a continuous distribution:

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

- Analogous to first moment (center of mass)

$$\int_{-\infty}^{\infty} x \cdot f(x) \cdot dx = \int_{0}^{180} x \cdot \frac{1}{180} \cdot dx = \frac{1}{180} \cdot \frac{x^2}{2}\Bigg|_{0}^{180} = \frac{180^2 - 0^2}{180 \cdot 2} = 90$$



Exercise: Mini-Yahtzee

- Previously studied the number of events per time:
  - Z: number of mini-Yahtzees in 1 minute
  - λ: average rate of mini-Yahtzee events

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

- Now study the time between adjacent events:
  - T: time between mini-Yahtzee events
  - λ: average rate of mini-Yahtzee events

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

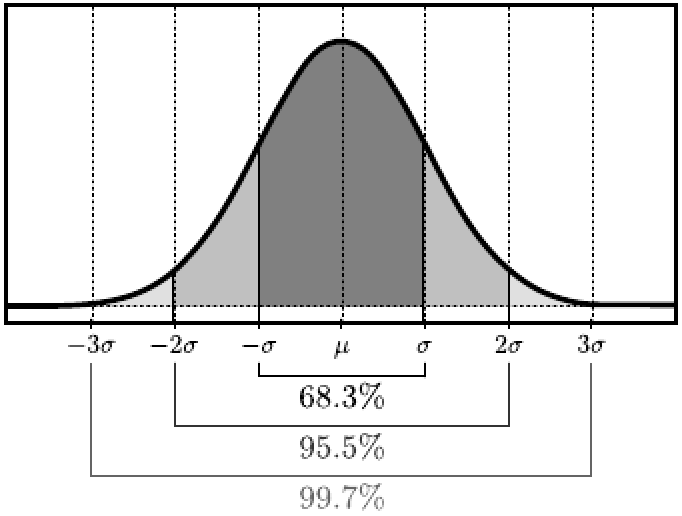
Continuous Distributions

- **Uniform:** Models a process with equally likely outcomes.
- **Triangular:** Models a process having minimum, maximum, and most likely values.
- **Normal:** Models a natural distribution with mean μ and standard deviation σ.
- **Chi-squared**: Models the sum of k squares of normally-distributed random variables.
- **Exponential:** Models the time between independent events with mean arrival rate λ (i.e. mean inter-arrival period $1/\lambda$).
- **Beta:** Widely-used general distribution.
- **Weibull:** Models the time until an event (e.g. component failure).
- **Lognormal:** Models the product of many independent random variables.

Normal Distribution

- Normal distribution (or Gaussian distribution) describes many natural systems

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \mathrm{X} \sim \mathrm{normal}(\mu, \sigma^2)$$



Confidence Intervals

Central Limit Theorem

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

- Assume n samples are randomly drawn from a population with mean μ and standard deviation σ
- Random variable of interest: sample mean $\bar{X}$

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

$$\mu_{\bar{x}} = E[\bar{X}] = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Dice Roller Sample Mean

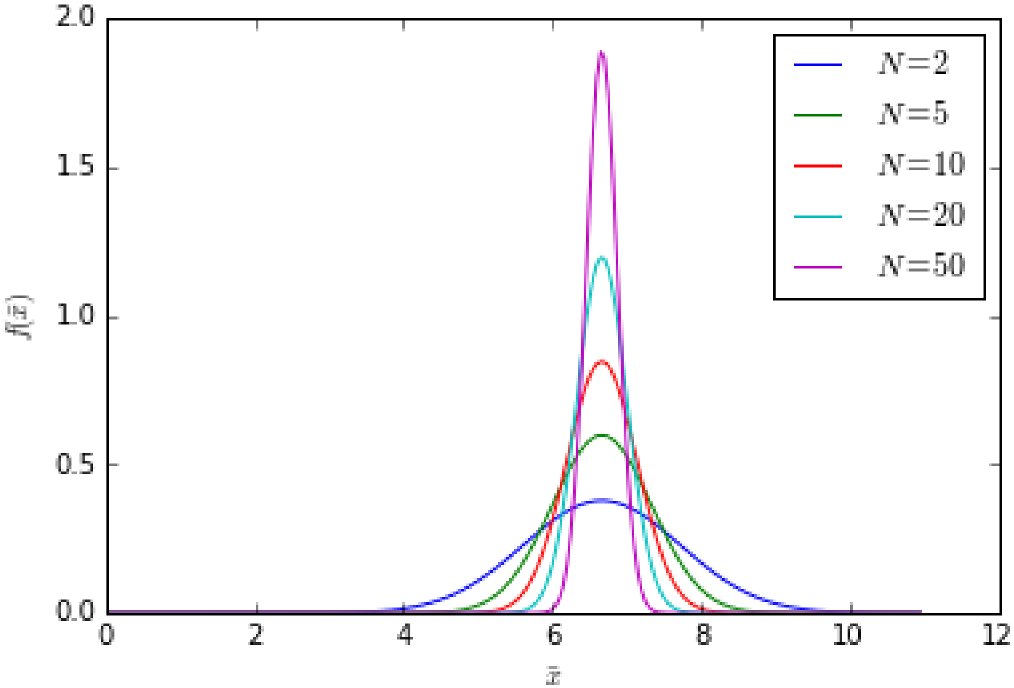
- $\mu_{\bar{x}} = E[\bar{X}] = \mu$

$$= n \cdot p = 10 \cdot \frac{2}{3} = 6.67$$

- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

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

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



Confidence Intervals

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

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

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

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

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

- Standard error of mean (SEM): $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

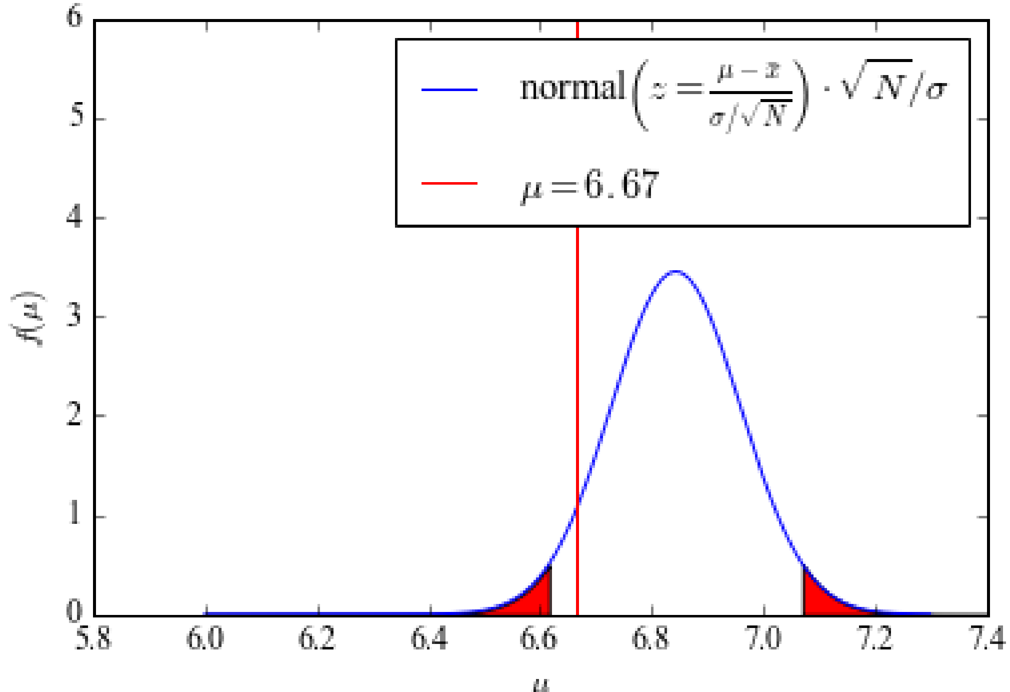
Estimating Population Mean

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

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

$$= [6.62, 7.07]$$

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



$\bar{x} - z_{0.975} \frac{s}{\sqrt{n}} = 6.62$

$\bar{x} + z_{0.975} \frac{s}{\sqrt{n}} = 7.07$