Probability

- Sets
  - Intersection of two sets(AND): $A \cap B$
  - Union of two sets(OR): $A \cup B$
- Defining Probability as a math function
  - $p(x)$ has some domain and some range [0,1]
  - Domain represents random outcomes – p assigns number to random outcomes
  - Random outcomes are explained via sets
    - Ex: In a sample of 10 people, what are the odds that average height > 6 feet?
- Frequency Theory of Probability
  - Ex: If a coin is tossed, what chance coin lands heads?
    - P(Head) – Probability of Heads
  - To find the probability – toss coin n = 100 times(sample size), find number of times heads appears, then divide by n
    - Sample proportion of heads in n = 100 tosses
  - Number of heads is random – random outcome
  - If number of heads is 60, proportion = 0.60
  - $\hat{p}$ = approximate probability – used for experiment of coin tosses, depends on tosses gained
  - Random Variable – depends on random outcomes of experiment
- Generalization of relative frequency theory:
  - Let E be some random event
  - Can estimate P(E) as $\hat{p} = \hat{p}(E) = \frac{\#\{E\}}{n}$
    - where we repeat the experiment n times **under identical conditions**
  - Theory only applies to events that can be repeated in the same conditions in the same way
  - $\lim\limits_{n \to \infty} \frac{\#\{E\}}{n}$ = P(E), as n increases, we get closer to the true probability
- Defining Probability(P) as a math function
  - Random Experiment – experiment with random outcomes
  - Sample Space(S) – set of all possible outcomes in an experiment
    - e.g. one coin toss: S = {H, T}
  - Cardinality – Number of elements in the set
    - if A is the set, notated by |A|
  - An event from a random experiment is a subset of the sample space
    - $E \subseteq S$, an event(E) can be a subset of the set of S
  - Two Sets, A and B are mutually exclusive when $A \wedge B$ is empty
    - No common elements

        - Three Sets(A_1,A_2,A_3) are pairwise mutually exclusive when A_1 and A_2 are mutually exclusive, A_1 and A_3 are mutually exclusive, and A_2 and A_3 are mutually exclusive
            - Can extend to K sets and a countably infinite number of sets
        - Probability of a subset of sample space computed within a Range:
            - $E \subseteq S \rightarrow R$
            - $E \subseteq S \rightarrow [0, 1]$
- Axioms of Probability
    - A probability function, P, must satisfy:
        - Probability of any event must be greater than or equal to 0
        - Probability of the entire sample space must be 1
            - Probability that some event in sample space happens is 1
        - Additivity(Finite)
            - Let E_1, E_2, and E_3 be 3 mutually exclusive events
            - P(E_1 or E_2 or E_3) = P(E_1) + P(E_2) + P(E_3)
            - Extends to countably infinite number of sets
- Probability Rules derived from Axioms
    - Complement Rule:
        - Let E be a set, the complement of E is $\overline{E} \equiv E' \equiv E^C$ is all elements outside of E but in S
        - Therefore $P(\overline{E}) = 1 - P(E)$
    - Probability of any event is bounded between 0 and 1
    - Inclusion Exclusion Probability:
        - P(E_1 U E_2) = P(E_1) + P(E_2) - P(E_1 and E_2)
            - Does not have to be mutually exclusive
            - Extends to k events – subtract all combinations to overlap

Counting Methods

- Multiplication:
    - 3 shirts, 4 pants – how many combinations?
    - Total number of ways: 3 shirts * 4 pants = 12 combinations
    - Two objects with n_1 and n_2 elements respectively, n_1 * n_2 ways to combine
    - Extends to k objects:
        - k objects with n_1,n_2,n_3,...,n_k elements respectively have n_1*n_2*n_3*...*n_k composites of k
- Permutations:
    - Interest in ordered items
    - 4 seats in a row with 4 persons
        - How many ways to arrange the 4 persons in the 4 seats:
            - s_1: 4 possibilities, s_2: 3 possibilities,...
            - So total ways: 4 * 3 * 2 * 1 = 24

    - 3 seats and 5 people
        - s_1: 5 possibilities, s_2: 4, s_3:3 – 5 8 4 * 3 = 60
        - or $\frac{5!}{2!} = \frac{5!}{(5-3)!}$
        - Generalized: Permutation of n objects taken k at a time:
            - nPk = $\frac{n!}{(n-k)!}$
- Combinations:
    - Interest in Unordered Objects
    - 9 people, 4 seats, order does not matter
    - s_1: 9, s_2: 8, s_3: 7, s_4: 6, with 4! possible ways to shuffle around current combinations
    - Alternate Expression: $\frac{9!}{(9-4)! * 4!}$
    - Generally: nCk = $\frac{n!}{(n-k)!k!}$, where k <= n

Hypergeometric Distribution

- Given N_1 numbers of class A and N_2 numbers of class B in a population
    - Population = N = $N_1 + N_2$
- Let X be the number of successes in a sample of size n
- P(X = x) = $\frac{(N_1, x)(N_2, n-x)}{(N, n)}$ – Hypergeometric Distribution
- $E(X) = \frac{N_1}{N_1+N_2}n$
- $Var(X) = n * \frac{n_1}{N} * \frac{N_2}{N} * \frac{N-N_1}{N-1}$

Conditional Probability

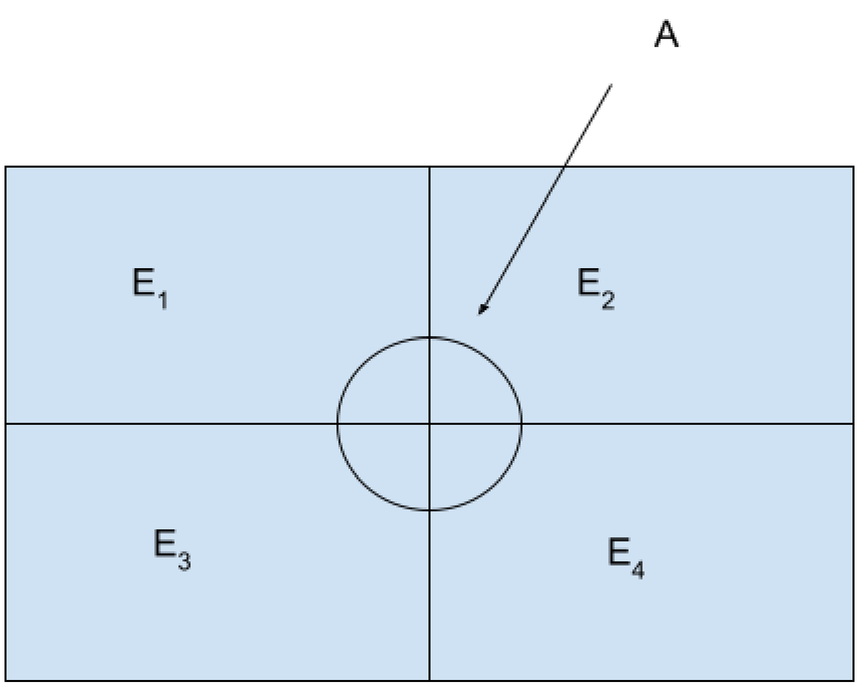
- Given information about outcomes
- Conditioning restricts us to a subset of the sample space
- Conditional Probability of A given that event B has happened is P(A | B)
- Multiplication Rule:
    - $P(A \cap B) = P(A|B) * P(B)$
    - $P(B \cap A) = P(B|A) * P(A)$
    - Extends of k events:
        - $P(A \cap B \cap C) = P(C|A\ and\ B) * P(B|A) * P(A)$

Independence

- Two events, A and B are statistically independent when$P(A \cap B) = P(A) * P(B)$
- De Morgan's Law:
    - $(A \cap B)^C = A^C \cup B^C$
    - $(A \cup B)^C = A^C \cap B^C$
- Let A be independent(⊥) of B, therefore
    - $A \perp B^C$

- $A^C \perp B$
- $A^C \perp B^C$

Law of Total Probability

- S:
    - 

- E_1, E_2, E_3 and E_4 are Pairwise Mutually Exclusive and
- $E_1 \cup E_2 \cup E_3 \cup E_4 = S$
    - Since both are true, considered a "partition" of S
- $P(A) = P([A \cap E_1] \cup [A \cap E_2] \cup [A \cap E_3] \cup [A \cap E_4]) =$
    - $P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) + P(A \cap E_4) =$

- $P(A \mid E_1) * P(E_1) + P(A \mid E_2) * P(E_2) + P(A \mid E_3) * P(E_3) + P(A \mid E_4) * P(E_4)$
- Extends to any K sets which form partitions of S

Bayes(Baby) Theorem:

- A, B are events in S
- $P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$, when P(B) > 0

Random Variables

- E = random event
- S = sample space
- P: $E \subseteq S \rightarrow [0, 1]$
    - Similarly: X, a random variable:
        - X: S → R
- Ex: Toss a coin 3 times, |S| = 8
    - Let X be the random variable defined by the number of tails in the outcome
    - Can we find a formula for P(X = x) where $x \in \{0, 1, 2, 3\}$?
- Types of Random Variables:
    - Discrete:
        - Finite: Number of tails in 3 tosses
        - Toss coin until first tail is observed – find number of tosses needed
    - Continuous:
        - Interval(possibly infinite of real line)
            - Ex: Let X be a random number in (0,1)
            - Ex: Let X be the lifetime of a certain battery
                - S_X = [0, infinity]

Distributions of Discrete Random Variables

- Let X be a **discrete** random variable
- Let f(x) = P(X = x) – probability mass function(PMF)
    - 0 < P(X = x) < 1
    - Sum over all probabilities equals 1
        - $\sum_{x \in S_x} P(X = x) = 1$
- Cumulative distribution function(CDF):
    - $F(x) = P(X \leq x)$
- Discrete Uniform Distribution:
    - $S_x = \{x_1, x_2, x_3, \ldots x_k\}$
        - $k \in \aleph = \{1, 2, 3, 4, \ldots\}$
    - PMF of X: $P(X = x) = \frac{1}{k}, x \in S_x$
    - CDF: $P(X \leq x) = \sum_{S_x}^{x} P(X = x)$

- $E(X) = \frac{k+1}{2}$
- $Var(X) = \frac{k^2-1}{12}$

- Bernoulli Distribution:
  - Probability of achieving a success from a Bernoulli Trial
    - Event with only two possible outcomes
  - $S_x = \{0,1\}$
  - $p = P(X = 1)$, $1 - p = q = P(X = 0)$
  - PMF: $p^x(1 - p)^{1-x}$
  - $E(X) = p$
  - $Var(X) = p(1 - p)$
- Geometric Distribution:
  - Independent Trials
    - Each trial observes a “success” or “failure”
    - Each trial, P = P(Success) and 1 - P = P(Failure) are the same, don’t depend on each other or the number of trials
  - Sequence of independent Bernoulli Trials
  - Let X = trial at which we get the first Failure
    - $S_x = \{1,2,3,...\}$
    - PMF = $P(X=x) = p^{x-1}(1-p)$
  - Let Y = trial until first success
    - PMF = $P(Y = k) = (1-p)^{k-1}p$
  - CDF: $F(x) = 1 - (1 - p)^x$
  - Memory-less property of geometric distribution:
    - $P(X > j + i \mid X > i) = P(X > j)$
  - $E(X) = \frac{1}{p}$
  - $Var(X) = \frac{1-p}{p^2}$
- Binomial Distribution
  - $X \sim Bin(n,p)$
  - *n* independent trials
  - On each trial, get either success or failure
    - Same probability of successes on each trial
    - p = P(Success)
  - X = Number of success in n trials
  - $S_x = \{0,1,2,3,...n\}$
  - $P(X = x) = nCx * p^x(1 - p)^{n-x}$
    - *n* = number of trials
    - *p* = probability of success

        - x = number of successes
    - $E(X) = np$
    - $Var(X) = np(1 - p)$
- Negative Binomial Distribution
    - Sequence of Independent Bernoulli Trials with same probability of success
    - X = trial at which k^{th} success happens
    - $P(X = x) = (x - 1)C(k - 1) * p^k * (1 - p)^{x-k}$
        - $S_x = \{k, k+1, k+2, \ldots\}$
    - $E(X) = \frac{k}{p}$
    - $Var(X) = \frac{k(1-p)}{p^2}$
- Poisson Distribution
    - Random counts in space / time
        - Customers arriving at a checkout counter over a given interval
        - Car accidents happening at a given intersection over a certain time period
    - Counts in non-overlapping subintervals are independent
    - Poisson process is characterized by a parameter(λ), represents the average number of counts per unit(interval)
    - $p(1\ count\ in\ the\ subinterval\ of\ length\ h) = \lambda * h$
    - $P(\geq 2\ counts\ in\ a\ small\ subinterval) = 0$
    - PMF: $P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}$
        - $S_x = \{0, 1, 2, \ldots, \infty\}$
        - For any Poisson: $P(X = 0) = e^{-\lambda}$
    - λ can be changed for different intervals:
        - $\frac{\lambda_x}{u_x} = \frac{\lambda_y}{u_y}$
    - $E(X) = \lambda$
    - $Var(X) = \lambda$

Mathematical Expectation

- Expectation = Weighted Averaging
- Let X be a discrete random variable:
    - $\mu = E(X)$
    - $E(X) = \sum_{x \in S_x} xP(X = x)$
    - $E[g(x)] = \Sigma g(x)P(X = x)$
    - $E(c_1g_1(X) + c_2g_2(X)) = c_1E[g(x)] + c_2E[g(x)]$
        - Extends to k functions(can add continually)

        - $E(X^2) \neq E(X)^2$

Variance / Standard Deviation

- Tells how far away the x's are from the mean
- $Var(X) = E[(X - \mu)^2]$
- Standard Deviation = $\sqrt{Var(X)}$
- $Var(X) = E(X^2) - E(X)^2$

Characteristics of Mathematical Expectation and Variance

- Given Random Variable X, let Y = a+bX, find E(Y) and Var(Y)
    - $\mu_y = a + b\mu_x$
    - $Var(Y) = b^2 Var(X)$
        - $\sigma_y = |b|\sigma_x$
- Physically moving distribution does not change the Variance

Continuous Random Variables

- If X is continuous
    - S_x is a continuous interval or union of such intervals
    - e.g. [0,1) and (0, ∞)
- Definition: Let X be a continuous random variable, the probability distribution function of X such that
    - $f(x) > 0,\ x \in S_x$
    - $\int_{-\infty}^{\infty} f(x)dx = 1$
- With a cumulative distribution function being
    - $\int_a^b f(x)dx$

Continuous Uniform Distribution

- PDF = $\frac{1}{b-a},\ when\ x \in (a,\ b)\ and\ 0\ when\ x\ not \in (a,\ b)$
    - where *a* and *b* are parameters
- CDF = $\frac{x-a}{b-a}$, where x is in the support of X

Exponential Distribution

- Similar to Poisson, where X is the arrival time of the 1st occurrence(continuous)
- P(Y < y) = 1 - P(Y > y)
    - = 1-P(X = 0 in (0, y))
    - = $1 - e^{-\lambda y}$
- PDF = $\frac{d}{dy} 1 - e^{-\lambda y} = \lambda e^{-\lambda y}$

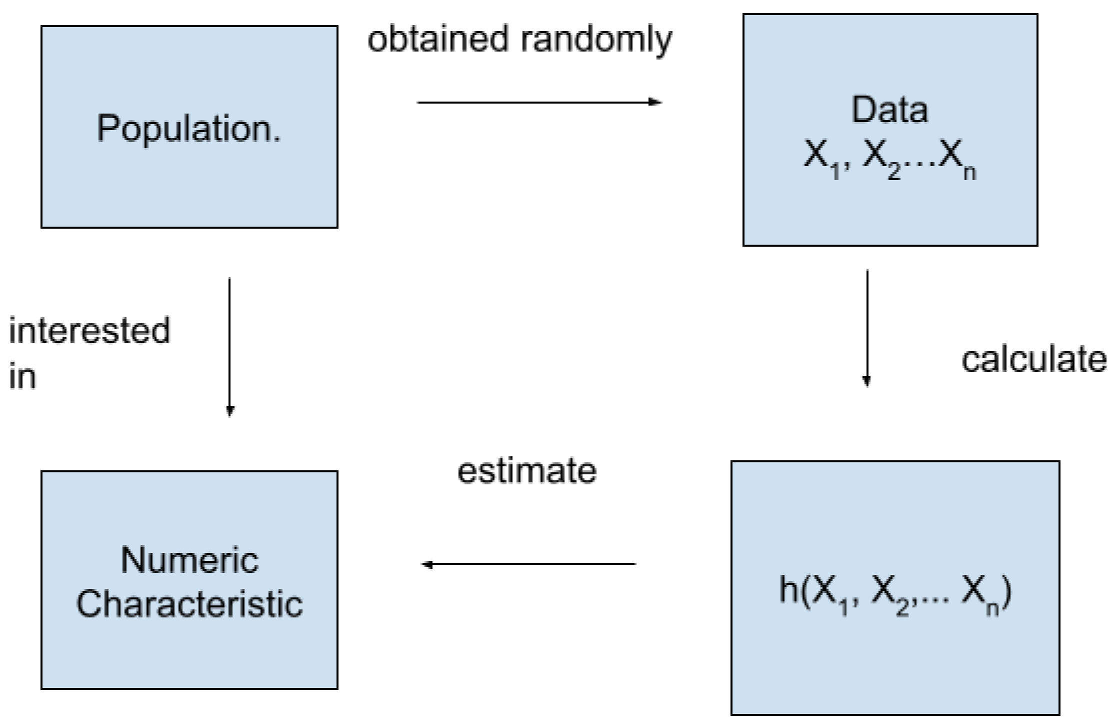
Expectation and Variance

- $E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx,$
    - where f(x) is the pdf
- $Var(X) = E(X^2) - [E(X)]^2$

Normal / Gaussian Distribution

- X ~ Normal(a, b^2)
- $PDF = f(x) = \frac{1}{\sqrt{2\pi b^2}} e^{\frac{-1}{2}(x-a)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2} * \frac{(x-\mu)^2}{2}}$
- E(X) = $\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi b^2}} e^{\frac{-1}{2}(x-a)^2} dx$ = a = μ
- Variance = $b^2 = \sigma^2$
- A standard Normal Distribution occurs when μ = 0 and $\sigma^2 = 1$
- SD Rule of Thumb:
    - Within one standard deviation of the mean – 68%
    - Within two standard deviations of the mean – 95%
    - Within three standard distributions of the mean – 99.6%
- Z-Score → $Y = \frac{X-\mu}{\sigma}$, number of standard deviations from mean
    - Y is distributed as a standard normal

Multiple Random Variables

- Taking *n* observations from a population
- Statistical Inference:
    - 

- A random sample of size *n* is defined such that all samples of n are equally likely
- Let $X_1, \ldots X_n$ be a sample
  - The X's are said to be independent and identically distributed(iid) when the X's are independent and also come from the same distribution

Joint Distribution of n random variables

- For Discrete
  - Joint = $P(X_1 = x_1$ and $X_2 = x_2 \ldots$ and $X_n = x_n)$
  - When independent:
    - Joint = $P(X_1=x_1) * P(X_2 = x_2) * \ldots * P(X_n = x_n)$
- For Continuous
  - Let X_1, X_2 be continuous random variables
    - $P(a < x_1 < b,\ c < x_2 < d) = \int_a^b \int_c^d f(x_1 x_2) dx_1 dx_2$
- Properties of Multiple Random Variables(independent or dependent, continuous or discrete)
  - $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$
  - Let X_1 and X_2 be independent random variables
    - $Var(aX_1 + bX_2) = a^2Var(X_1) + b^2Var(X_2)$
  - Generally:
    - $Var(g(X_1) + h(X_2)) = Var(g(X_1)) + Var(h(X_2))$

Sampling

- If X's are taken with replacement sampling from the same population, the X's are iid
- If X's are taken without replacement sampling from the same population, the X's are dependent but identically distributed
  - However, if the population is substantially larger than the sample size, the draws are "effectively" independent

Expectation and Variance for Multiple Random Variables

- $E(a_1X_1 + \ldots a_nX_n) = a_1E(X_1) + \ldots + a_nE(X_n)$
- $Var(a_1X_1 + \ldots a_nX_n) = a_1^2\sigma^2_1 + \ldots + a_n^2\sigma_n^2$ if independent
  - $E(g_1(X_1) + g_2(X_2) + \ldots g_n(X_n)) = E[\sum_{i=1}^{n} g_i(X_i)] = \sum_{i=1}^{n} E[g_i X_i]$
  - $V[\sum_{i=1}^{n} g_i(X_i)] = \sum_{i=1}^{n} V[g_i X_i]$, when independent
  - $E[\prod_{i=1}^{n} g_i(X_i)] = \prod_{i=1}^{n} E[g_i X_i]$, when independent

Moment Generating Functions

- Can be used to find distribution of same statistics
- Every distribution has unique MGF
  - The MGF of X is $M_x(t) = E(e^{tx})$

- Used for:
  - Finding Moments of X
    - The kth moment about the number b is $E[(X-b)^k]$,
      - if $b = \mu_x$, k = 2, Variance occurs
    - The kth derivative of of $M(t) = E(X^k)$
  - Finding distributions of multiple random variables
    - Let X_1 and X_2 be independent random variables and $S = X_1 + X_2$
      - $M_s(t) = E[e^{tX} * e^{tX}] = M_X(t) * M_X(t)$, extends to n random variables

Gamma Distribution

- Gamma function, extension of n!
  - $\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy$, where t >= 0
  - $\Gamma(n) = (n - 1)!$, where n is a positive integer
- Gamma PDF:
  - $f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} * x^{\alpha-1} * e^{\frac{-x}{\theta}}, x > 0, \alpha > 0, \theta > 0$
- $E(X) = \alpha\theta$
- $Var(X) = \alpha\theta^2$
- α = shape parameter
- θ = scale parameter

Chi-Square(χ^2) Distribution

- special Gamma with $\alpha = \frac{r}{2}$, $\theta = 2$, where r > 0
- $\chi^2(r) = Gamma(\frac{r}{2}, 2)$
- $f(x) = \frac{1}{\Gamma(\frac{r}{2})2^{\frac{r}{2}}} * x^{\frac{r}{2}-1} * e^{\frac{-x}{2}}, x > 0$, r → degrees of freedom
- $E(X) = r$
- $Var(X) = 2r$
- As $r \to \infty$, $\chi^2(r) \to N(r, 2r)$

Central Limit Theorem

- $X_1,\ldots,X_N$ independent; $L = a_1X_1 + \ldots a_nX_n$, where a_i is a real number
- if X's are normal, L is exactly $N(\mu_L, \sigma_L^2)$
- if X's are not normal, but n is large enough(> 29)
  - $L \sim N(\mu_L, \sigma_L^2)$, approximately
  - P(L < a) is approximately P(Normal < a)
- Normal Approximation with a continuity correction(discrete to continuous)
  - $P(a < Y < b) = P(a + 0.5 \le Y \le b - 0.5)$

Maximum Likelihood Estimation

- Finding parameters which maximize the likelihood of the data occurring
- 1. $l(p) = L(p \mid data)$
- 2. Let $L(p) = ln(l(p))$
- 3. Find $\frac{dL}{dp} = 0$, solve for p
- 4. Check if p is a max(don't have to for STAT 315)
- Likelihood Principle: Find $\theta(function\ of\ data)\ s.t.\ l(\theta)$ is largest

Terminology

- For observed data $X_1, \ldots, X_N \rightarrow \mu_{ML} = \overline{X}$
- How good of an estimator is $\overline{X}$ for μ?
  - For now, estimator ± SD(estimator)
  - where the SD of an estimator is the standard error

Confidence Interval

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2), \mu_{ML} = \overline{X},\ SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}$
- $X \pm SE(\overline{X})$, since X is normal, the odds that the value falls in this interval is 0.68
- Probability that random interval captures μ is 68%, not the probability that μ falls in interval
- a% confidence interval → If I collect many data sets, computing an interval for each dataset, a% of the intervals will capture the true value of μ
- "Confidence Factor" – coefficient which controls confidence – look up in chart

What makes a good estimator?

- Bias –
  - $Bias(\hat{p}) = E(\hat{p}) - p$, if Bias is 0, unbiased estimator
- Low Variance of $\hat{\theta}$, $V(\overline{X}) = \frac{\sigma^2}{n} --> 0\ as\ n -> \infty$

Confidence Interval for mean

- $\overline{X} \pm Z_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$, for known population standard deviation
- $\overline{X} \pm t_{\frac{\alpha}{2}}(\frac{S}{\sqrt{n}})$, for unknown population standard deviation

Confidence Interval for Difference of Means

- $\overline{X} - \overline{Y} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$, for known population standard deviation
- If unknown variances, homogeneity assumption of $\sigma_x^2 = \sigma_y^2$
  - $\overline{X} - \overline{Y} \pm t_{\frac{\alpha}{2}}(n + m - 2)(S_p)\sqrt{\frac{1}{n} + \frac{1}{m}}$,
    - where $S_p^2 = \frac{n}{n+m}S_x + \frac{m}{n+m}S_y$

Confidence Interval for Proportions

- $\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, for large n by CLT
- $\hat{p}_1 - \hat{p}_2 \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$

Hypothesis Testing

- Null Hypothesis: H_o → nothing happens
- Alternative Hypothesis: H_1 → something happens
- Steps:
    - Specify H_0 and H_1
    - Test statistic, "compare" data to H_o
    - statistics → E(stat | H_o)
    - Find p-value placing stat – E(stat | H_o) on a probability
    - Make a decision based on alpha level
- If p value < alpha level, reject null hypothesis
- If p value > alpha level, fail to reject null hypothesis
- p-value does not equal probability that null hypothesis is true
- If p-value is very small, does not prove H_o is false
- Any test of hypothesis can be made to reject null hypothesis with a large enough sample size
    - For Confidence Intervals, just makes the interval smaller, no real issues
- Tests are for parameters, not statistics