Statistics

Set Theory Definitions:
- Intersection: $A \cap B$ (A and B)
- Union: $A \cup B$ (A or B)
- Complement: $A^C$ (not A)
- Disjoint: $A \cap B = \emptyset \iff A$ and B are disjoint

Set Theory Thoughts:
- $A \cup A^C = \Omega$ (the entire set)
- $A \cap A^C = \emptyset$
- Commutative Law:
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$
- Associative Law:
  - $(A \cap B) \cap C = A \cap (B \cap C)$
  - $(A \cup B) \cup C = A \cup (B \cup C)$
- Distributive Law:
  - $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
  - $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Probability

Sample Space ($\Omega$) = the set of all possible outcomes; this depends on what is actually being measured.

Event = a subset of the sample space

Probability (P) is a function that maps subsets of $\Omega$ to real numbers and satisfies the following axioms:
- 1. $P(\Omega) = 1$
- 2. If $A \subseteq \Omega$ then $P(A) \geq 0$
- 3. If $A_1$ and $A_2$ are disjoint, then $P(A_1) + P(A_2) = P(A_1 \cup A_2)$

Conditional Property: The probability of one event given another:

$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

If $P(A) = P(A \mid B)$ then A and B are statistically independent.

Other properties of P:
- $P(A^C) = 1 - P(A)$
- $P(\emptyset) = 0$
- If $A \subseteq B$ then $P(A) \leq P(B)$
- $P(A \cap B) = P(A) P(B \mid A) = P(B) P(A \mid B)$
  - if A and B are statistically independent
- $P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)} = \frac{P(A) + P(B \mid A^C) P(A^C)}{P(B \mid A^C) P(A^C)}$ [Bayes’ Rule]

Other method of finding joint probabilities: a tree
- Split $\Omega$ into A and $A^C$ and label each edge with the probability.
- Split each leaf into B and $B^C$ and label each edge with $P(B \mid$ previous conditions)
- Multiplying the numbers to the leaf gives the probability of that intersection of events

If every outcome of $\Omega$ is equally likely, then $P(A) = |A| / |\Omega| = $ the fraction of outcomes in A

Counting Theory

Permutation = an ordered arrangement

Permutations of K of N cards is $N!/(N-K)!$

Combination = an unordered arrangement

Combinations of k of n cards is “n choose k” = $n!/(k!(n-k)!)$

Then, multiply…

Random Variables

Random Variable = a variable whose value is determined by a random process

Random Variables may be discrete or continuous

Cumulative Distribution Function: $F_X(x) = P(X \leq x)$

Necessary Properties

$\lim (x \to -\infty) F_X(x) = 0$
lim (x → ∞) F_X(x) = 1
F_X(x) is non-decreasing.

The distance between lim (x → a') and lim (x → a') is P(X = a)

Probability Density Function: f_X(x) = d F_X(x) / dx (where this is defined)

Necessary Properties
f_X(x) ≥ 0 for all x
∫ f_X(x) dx = 1 [from -∞ to ∞]

Joint Density and Cumulative Distribution Functions: F_{XY}(x, y) = P(X ≤ x and Y ≤ y)
F_{XY}(x, y) = \sum P(X = u and Y = v) = \int \int f_{xy}(u, v) dv du
f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} (F_{XY}(x, y)) / \partial x \partial y
As before, f_{XY}(x, y) ≥ 0 for all (x, y) and \int f_{XY}(x, y) dx dy = 1 (from -∞ to ∞ for both)
P(X, Y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx \, dy.
Marginal Densities: f_X(x) = \int f_{XY}(x, y) dy [from -∞ to ∞]
Marginal Densities: f_Y(y) = \int f_{XY}(x, y) dx [from -∞ to ∞]

If X and Y are independent, then F_{XY}(x, y) = F_X(x) F_Y(y) and f_{XY}(x, y) = f_X(x) f_Y(y).
Conditional Distribution: P(X = x | Y = y) = P(X = x, Y = y) / P(Y = y)

Expected Value (mean) = the weighted average over all values of a random variable
E(X) = \sum x \cdot P(X=x) [over all possible x]
E(X) = \int x \cdot f_X(x) \, dx [from -∞ to ∞]
This is provided that \int |x| f_X(x) dx or \sum |x| \cdot P(X = x) does not diverge. Otherwise, E(X) is undefined.
E(g(X)) = \int g(x) \cdot f_X(x) \, dx [Law of the Unconscious Statistician]
E(g(X, Y)) = \int \int g(x, y) \cdot f_{XY}(x, y) \, dx \, dy.
Conditional Mean: \mu_{Y|X} = E(Y | X = x) = \int y \cdot f_{Y|x}(y | x) \, dy [where x is fixed and f_{Y|x}(y | x) depends on x]

Variance (\sigma^2) = the measure of the spread of the values
Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2
\sigma_{X|Y}^2 = Var(Y | X = x) = E((Y - \mu_{Y|X})^2 | X = x) = E(Y^2 | X = x) - (E(Y | X = x))^2
Law of Total Variance: Var(Y) = Var(Y | X) + Var(E(Y | X))

Standard Deviation (\sigma) = the square root of the variance

Covariance: How Variables Move Together
Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y).
Cov(X, Y) = 0 if X and Y are independent (not vice versa). X and Y are uncorrelated if Cov(X, Y) = 0.
Cov(X, X) = Var(X).
Cov(a + \sum b_i X_i; c + \sum d_i Y_j) = \sum \sum b_i d_j Cov(X_i; Y_j)
Var(X + Y) = Var(X) + Var(Y) + 2 Cov(X, Y).

Correlation: \rho_{XY} = Cov(X, Y) / \sqrt{(Var(X)Var(Y))}
-1 ≤ \rho_{XY} ≤ 1
|\rho_{XY}| = 1 \iff Y = a + bX.

Transformations of Random Variables (CDF method): Suppose X is a random variable and Y = g(X).
1. Express F_Y(y) in terms of F_X(x).
   F_Y(y) = P(Y ≤ y) = P(g(X) ≤ y)
2. Differentiate to find f_Y(y),
   f_Y(y) = d [F_Y(y)] / dy (Hint: Use the chain rule and f_X(x).)
If g^{-1}(x) exists, then
   f_Y(y) = \frac{d (g^{-1}(y))}{dy} (f_X (g^{-1}(y)))

Functions of Jointly Distributed Random Variables:
F_S(s) = P(g(X, Y) < s) = \int_{g_X^{-1}(s)}^{g_Y^{-1}(s)} f_{XY}(x, y) \, dx \, dy.

Moments: E((E - E(X))^k) is the kth central moment.
The moment generating function is M_k(t) = E(e^{tX}) = \sum e^{tx} \cdot P(X = x) = \int e^{tx} \cdot f_X(x) \, dx.
If M_k(t) exists in an open interval about 0, M_k(0) is the kth moment of X (f^k is the kth derivative).
If M_X(t) exists in an open interval about 0 then M_X(t) uniquely determines the distribution of X.
If Y = a + bX then M_Y(t) = e^{at} M_X(bt).
If X and Y are independent and S = X + Y, then M_S(t) = M_X(t) M_Y(t).

Statistics
Statistic: A numerical summary of data. It is a function of the random variables that are being measured.
(Therefore, they are also random variables.)
- The usefulness of statistics can be measured by:
  - bias(θ-hat) = E(θ-hat) - θ = expected distance from the mean. A statistic is unbiased if the expectation of the statistic is the true value for the population.
  - Var(θ-hat) = E((θ-hat - E(θ-hat))^2)
  - Mean squared error: MSE = E((θ-hat - θ)^2) = Var(θ-hat) + bias(θ-hat).
  - Asymptotically unbiased. (lim (n → ∞) E(θ-hat) = θ)
  - Consistent: θ-hat approaches θ in probability. (ie: lim (n → ∞) P(|θ-hat - θ| > ε) = 0.)

Simple Random Sample (SRS): A sample from a population of size N is a simple random sample of size n if every set of n (usually distinct) individuals is equally likely to be the chosen sample.

- If we take an SRS of size n from a population of size N and measure \( X_i \sim [\mu, \sigma^2] \) then X-bar = \( \sum X_i/n \sim [\mu, (\sigma^2/n)(1 - (n-1)/(N-1))] \). (This is an unbiased statistic.)
- Note that the population mean and variance are the same as the mean and variance of the distribution of each \( X_i \).
- (Central Limit Theorem.) \( \lim (n \rightarrow \infty) P(\text{X-bar} - \mu, \sigma^2/n < z) = \Phi(z) \). (The distribution of X-bar approaches N(μ, σ^2/n).)
- (Law of Large Numbers.) With infinite N, \( \lim (n \rightarrow \infty) P(|\text{X-bar} - \mu| > t) = 0 \) for all t > 0. (The probability of the sample mean being any finite distance from the population mean goes to zero as the sample size gets infinite.)
- If \( X_i \) and \( X_j \) are two elements of an SRS then \( \text{Cov}(X_i, X_j) = -\sigma^2/(N-1) \). (Big population independent; with replacement → independent.)

Estimating Variance: (Not as easy.)

- \( \sigma^2 = \sum (X_i - \text{X-bar})^2/n \) [biased but consistent]
- \( s^2 = \sum (X_i - \text{X-bar})^2/(n-1) \) [larger mean squared error]
- If this is the variance of a binomial distribution, use the variance if p = .5. (This is the maximum possible variance.)
- If \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \) then \( \sum (X_i - \mu)/\sigma^2 \sim \chi^2_{n-1} \); in other words, \( (n-1)s^2/\sigma^2 \sim \chi^2_{n-1} \).
  - This depends on the fact that \( s^2 \) and X-bar are independent random variables.
  - This leads to confidence intervals for the variance, as well.

Confidence Intervals: Intervals which, with certain confidence, contain the true mean.

- These are constructed to a probability interval for X-bar. Only now X-bar is known and \( \mu_X \) might not be. Since \( \mu \) is not a random variable, however, this is not a probability interval.
- X-bar ± \( z*\sigma_{x-bar} \) is a confidence interval with \( z \) from N(0, 1) if \( \sigma_{x-bar} \) is known.
  - \( \sigma_{x-bar} = (\sigma^2/n)(\sqrt{1-(n-1)/(N-1)}) \)
  - If \( \sigma \) is not known, we estimate the standard error (see estimating variance above).
  - We may also use the T-distribution (with n-1 degrees of freedom) instead – this takes the uncertainty of variance into account.

- Intervals get smaller with less confidence, more elements in the sample, or a smaller population variance.
- The \( z*\sigma_{x-bar} \) is the margin of error.

Estimating Parameters:

- Method of Moments
  - Let \( \mu_k = E(X^k) \); this is the \( k^{th} \) moment. We may estimate \( \mu_k \) by \( (\text{X-bar})_k = \sum X^k/n \). This is an unbiased estimate.
  - Express parameters as function of \( \mu_1, \mu_2, \ldots \) Then use the estimates of \( \mu_k \) to estimate the parameters. (\( \Theta_i \) = \( f_1(\mu_1, \mu_2, \ldots) \), and \( \Theta_i\)-hat = \( f_i(\text{X-bar}_1, \text{X-bar}_2, \ldots) \).)
  - If the \( f_i \) are continuous, then \( \Theta \)-hat is a consistent estimator for \( \Theta \).

- Maximum Likelihood Estimate
  - Suppose \( X_1, \ldots, X_n \sim f(x \mid \Theta) \). (The distribution depends on the values of some parameters.)
  - \( L(\Theta) = \prod f(x \mid \Theta) \). This is the likelihood of getting this distribution of \( X_i \) given a value of the parameter. The \( \Theta \) that maximizes this is the maximum likelihood estimate.
  - The log-likelihood function is: \( l(\Theta) = \ln(L(\Theta)) = \sum \ln(f(x, \mid \Theta)) \). Maximizing this is equivalent to maximizing L(\( \Theta \)) and is often easier.
  - The information function is: \( I(\Theta) = E(\partial^2 \ln(f_i(x \mid \Theta)) / \partial \Theta^2) = -E(\partial^2 \ln(f_i(x \mid \Theta)) / \partial \Theta^2) \)
  - The variance of \( \Theta \)-hat is \( 1/nI(\Theta) \); this may be estimated with \( \Theta \)-hat.
- Maximum likelihood estimates are asymptotically unbiased, consistent and asymptotically normal. If \( \theta \) is the mle of \( \theta_0 \) then \( \theta \sim N(\theta_0, \frac{1}{n}I(\theta_0)) \) as \( n \) grows.
- Cramer-Rao Lower Bound: Let \( T(X_1, \ldots, X_n) \) be an unbiased estimator for \( \theta \). Then \( \text{Var}(T(X_1, \ldots, X_n)) \geq \frac{1}{n}I(\theta) \). (This is the best possible variance.)
- For multiple parameters, the maximum likelihood estimates maximize \( l(\theta) \) for all.
  (Take all partials and solve.)
- Transformation Invariance: If \( \theta \) is the mle for \( \theta_0 \) and \( g \) is an invertible function then the mle for \( \phi = g(\theta_0) \) is \( g(\theta) \).
- Likelihood Principle. Two likelihood functions are equivalent if \( L(\theta) / L'(\theta) \) is constant (or if \( l(\theta) = c + l'(\theta) \)). If two experiments yield equivalent likelihood functions then the inference about \( \theta \) made from each should be the same.
- Sufficient Statistics
  - A statistic \( T(X_1, \ldots, X_n) \) is sufficient for \( \theta \) if the conditional distribution of \( X_1, \ldots, X_n \) given \( T(X_1, \ldots, X_n) \) does not depend on \( \theta \).
  - Equivalently, \( P(X_1 = x_1, \ldots, X_n = x_n \text{ and } T = t) / P(T = t) \) is not a function of \( \theta \).
  - Factorization Theorem. \( T(X_1, \ldots, X_n) \) is sufficient for \( \theta \) if and only if \( f(x_1, \ldots, x_n | \theta) = g(T(x_1, \ldots, x_n), \theta) h(x_1, \ldots, x_n) \).
  - A sufficient statistic, \( T \), is minimal if for any other sufficient statistics \( T' \), \( T \) is function of \( T' \).
- Exponential Family of Probability Densities: Densities which can be described by \( f_X(x|\theta) = \exp\left(\sum c_j(\theta)T(x) + d(\theta) + S(x)\right)\delta(x \in A) \), where \( A \) does not depend on \( \theta \). In this case, \( \sum T_j(X_i) \) for each \( j \), are the sufficient statistics for \( f_x \).
- If \( T(X) \) is sufficient for \( \theta \) then the maximum likelihood estimator of \( \theta \) is a function of \( T(X) \).

Neyman-Pearson Paradigm: Choosing between two hypotheses, \( H_0 \) and \( H_a \).
- Specify null and alternative hypotheses about the underlying data.
- Specify a rejection region for \( T(X) \), some function of the data. If \( T(X) \) is in the rejection region, then \( H_0 \) is rejected. Otherwise, it is not rejected.
  - The significance level of the test, \( \alpha \), is the probability of \( T(X) \) being in the rejection region when \( H_0 \) is true (a type I error).
  - The power of the test, \( 1 - \beta \), the probability of failing to reject \( H_0 \) when it is false (type II error).
- By the Neyman-Pearson Lemma, the most powerful test at any fixed significance level is based on the likelihood ratio. (At least, for simple hypotheses, and in some general cases as well.) We reject when the likelihood ratio is small.
  - For two simple hypotheses, \( T(X) = L(\theta_0) / L(\theta_a) \).
  - Generalized Likelihood Ratio, when \( H_0 \) is \( \theta \in \omega_0 \) and \( H_a \) is \( \theta \in \omega_a \):
    - \( \Lambda^* = \max (\theta \in \omega_0) L(\theta) / \max (\theta \in \omega_a) L(\theta) \)
    - \( \Lambda = \max (\theta \in \omega_0 \cup \omega_a) L(\theta) / \max (\theta \in \omega_0 \cup \omega_a) L(\theta) \)

Distributions (Discrete)

Discrete Uniform: Each of \( n \) outcomes has an equally likely outcome.

\[
P(X = x) = \frac{1}{n} \quad \text{if } x \text{ is a possible outcome}
0 \quad \text{otherwise}
\]

Bernoulli: \( X \) is the number of successes in a single trial; \( X = \{0, 1\} \)

\[
P(X = x) = \begin{cases} 
1 - p & \text{if } X = 0 \\
p & \text{if } X = 1 \\
0 & \text{otherwise}
\end{cases}
\]

\[E(X) = p\]

Binomial Random Variable: the sum of \( n \) independent Bernoulli random variables with equal probability.

Equivalent to sampling with replacement \( n \) times from a population with proportion \( p \) successes.

\[
P(X = x) = p^x (1-p)^{n-x} \quad (n \text{ choose } x)
\]

\[E(X) = np\]
Geometric: If \( X \) is the number of trials up to and including the first success (which has probability \( p \)), then \( X \sim \text{Geometric}(p) \).

\[
P(X = x) = (1-p)^{x-1}p
\]

Negative Binomial: If \( X \) is the number of trials up to and including the \( r \)th success (which has probability \( p \)), then \( X \sim \text{Negative Binomial}(r, p) \).

\[
P(X = x) = (1-p)^{x-r}p^r \binom{x-1}{r-1}
\]

Hypergeometric: If \( X \) is the number of successes in a random sample without replacement of size \( n \) from a population of size \( N \), in which \( r \) of the \( N \) elements are successes, \( X \sim \text{Hypergeometric}(N, n, r) \).

\[
P(X = x) = \binom{r}{x} \binom{N-r}{n-x} / \binom{N}{n}
\]

if \( X = 0, 1, \ldots, \min\{n, r\} \)

Poisson: If \( X \) is the number of events that occur in a fixed interval, with average rate \( \lambda \), \( X \sim \text{Poisson}(\lambda) \).

Assumptions of a Poisson Process:

- No simultaneous events
- The number of events in any disjoint time intervals are independent.
- The probability distribution of the number of events in two intervals of the same length are the same.

\[
P(X = x) = \lambda^x e^{-\lambda} / x!
\]

for \( X = 0, 1, 2, \ldots \)

\[
E(X) = \text{Var}(x) = \lambda
\]

Continuous Uniform: if \( X \) has a uniform probability on \([a, b]\) then \( X \sim \text{Uniform}(a, b) \)

\[
f_X(x) = \begin{cases} 
1/(b-a) & \text{if } a \leq x \leq b \\
0 & \text{otherwise}
\end{cases}
\]

Exponential: \( X \sim \text{Exp}(\lambda) \) if

\[
f_X(x) = \begin{cases} 
0 & \text{if } x < 0 \\
\lambda e^{-\lambda x} & \text{if } x \geq 0
\end{cases}
\]

Interesting Note: This function is memory-less.

Gamma: \( X \sim \text{Gamma}(\alpha, \lambda) \) if

\[
f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}
\]

where \( x \geq 0 \)

\[
\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du \quad [\text{from 0 to } \infty]
\]

Note: \( \Gamma(k+1) = k \Gamma(k) \)

\[
E(X) = \alpha / \lambda
\]

Normal Distribution: If \( X \sim \text{N}(\mu, \sigma^2) \) then

\[
f_X(x) = \exp(- (x-\mu)^2 / 2\sigma^2) / \sqrt{2\pi}\sigma^2
\]

\[
M_X(t) = \exp(t\mu + \sigma^2t^2/2)
\]

Properties of the normal distribution:

- Centered at and symmetric about \( \mu \)
- Inflection points at \( \mu \pm \sigma \) (one standard deviation away)

Links Between Distributions:

- Bernoulli(p) \sim \text{Binomial}(1, p)
- Geometric(p) \sim \text{Negative Binomial}(1, p)
- The limit of the Hypergeometric as the population increases and the proportion stays constant is Binomial \((n, r / N)\).
- The limit of the Binomial as the number of samples \((n)\) increases and the probability \((p)\) approaches 0 in a way that \(n*p\) approaches a constant is Poisson\((np)\).
- The time between Poisson events is distributed as Exponential
- Exp(\(\lambda) \sim \text{Gamma}(1, \lambda)
- Standard Normal \sim \text{N}(0, 1)
- \(\chi^2(n) \sim \text{Gamma}(n/2, 1/2)\)
- \(\chi^2(n)\) is generated by finding the distribution of \(\sum(X_i - \mu_i)^2/\sigma_i^2\) when \(X_i \sim \text{N}(\mu_i, \sigma_i^2)\).
- If \(Z \sim \text{N}(0, 1)\) and \(U \sim \chi^2_k\) then \(Z / \sqrt{(U/k)} \sim t_k\) (Student’s t distribution).

Finding probabilities in the normal distribution (using a table with the standard distribution):

\[
P(X \leq x) = P(Z \leq (x - \mu)/\sigma)
\]