Statistics

Set Theory Definitions:

Intersection: $A \cap B$ (A and B)

Union: $A \cup B$ (A or B)

Complement: A^C (not A)

Disjoint: $A \cap B = \varnothing \Leftrightarrow$ A and B are disjoint

Set Theory Thoughts:

$A \cup A^C = \Omega$ (the entire set)

$A \cap A^C = \varnothing$

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 (Ω) = 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 Ω 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:

$P(A \mid B) = P(A \cap B) / P(B)$

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

Other properties of P:

$P(A^C) = 1 - P(A)$

$P(\varnothing) = 0$

If $A \subseteq B$ then $P(A) \leq P(B)$

$P(A \cap B) = P(A)\, P(B \mid A) = P(B)\, P(A \mid B)$

$= P(A)\, P(B)$ *if* A and B are statistically independent

$P(A \mid B) = P(B \mid A)\, P(A) / (P(B \mid A)\, P(A) + P(B \mid A^C)\, P(A^C))$ [Bayes' Rule]

A method of finding joint probabilities: a tree

- Split Ω 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 | previous conditions)
- Multiplying the numbers to the leaf gives the probability of that intersection of events

.

If every outcome of Ω 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

$\lim (x \rightarrow -\infty)\, F_X(x) = 0$

lim ($x \to \infty$) $F_X(x) = 1$

$F_X(x)$ is non-decreasing.

The distance between lim ($x \to a^+$) and lim ($x \to a^-$) is $P(X = a)$

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

Necessary Properties

$f_X(x) \geq 0$ for all x

$\int f_X(x)\, dx = 1$ [from $-\infty$ to ∞]

Joint Density and Cumulative Distribution Functions: $F_{XY}(x, y) = P(X \leq x \text{ and } Y \leq y)$

$F_{XY}(x, y) = \sum P(X = u \text{ and } Y = v) = \int^x \int^y f_{xy}(u, v)\, dv\, du$

$f_{XY}(x, y) = \partial^2 (F_{XY}(x, y)) / \partial x\, \partial y$

As before, $f_{XY}(x, y) \geq 0$ for all (x, y) and $\iint f_{XY}(x, y)\, dx\, dy = 1$ (from $-\infty$ to ∞ for both)

$P((X, Y) \in A) = \iint_A f_{XY}(x, y)\, dx\, dy$.

Marginal Densities: $f_X(x) = \int f_{XY}(x, y)\, dy$ [from $-\infty$ 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: $P(X = x \mid Y = y) = P(X = x, Y = y) / P(Y = y)$

Expected Value (mean) = the weighted average over all values of a random variable

$E(X) = \sum x\, P(X=x)$ [over all possible x]

$E(X) = \int x\, f_X(x)\, dx$ [from $-\infty$ to ∞]

This is provided that $\int |x|\, f_X(x)\, dx$ or $\sum |x|\, P(X = x)$ does not diverge. Otherwise, E(X) is undefined.

$E(g(X)) = \int g(x)\, f_X(x)\, dx$. [Law of the Unconscious Statistician]

$E(g(X, Y)) = \iint g(x, y)\, f_{xy}(x, y)\, dx\, dy$.

Conditional Mean: $\mu_{y \mid x} = E(Y \mid X = x) = \int y\, f_{y|x}(y \mid x)\, dy$ [where x is fixed and $f_{y|x}(y \mid x)$ depends on x]

Law of Total Expectation: $E(Y) = E(E(Y \mid X))$

Variance (σ^2) = the measure of the spread of the values

$Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$

$\sigma_{y|x}^2 = Var(Y \mid X = x) = E((Y - \mu_{y|x})^2 \mid X = x) = E(Y^2 \mid X = x) - (E(Y \mid X = x))^2$

Law of Total Variance: $Var(Y) = E(Var(Y \mid X)) + Var(E(Y \mid X))$

Standard Deviation (σ) = the square root of the variance

Covariance: How Variables Move Together

$\sigma_{XY} = Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$.

$Cov(X, Y) = 0$ if X and Y are independent (not vice versa). X and Y are <u>uncorrelated</u> if $Cov(X, Y) = 0$.

$Cov(X, X) = Var(X)$.

$Cov(a + \sum b_i X_x, c + \sum d_j Y_j) = \sum\sum b_i d_j\, Cov(X_i, Y_j)$

$Var(X + Y) = Var(X) + Var(Y) + 2\, Cov(X, Y)$.

Correlation: $\rho_{xy} = Cov(X, Y) / \sqrt{(Var(X)Var(Y))}$

$-1 \leq \rho_{XY} \leq 1$

$|\rho_{XY}| = 1 \Leftrightarrow Y = a + bX$.

Transformations of Random Variables (CDF method): Suppose X is a random variable and Y = g(X).

1. Express $F_Y(y)$ in terms of $F_X(x)$.

$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$

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

$f_Y(y) = (d (g^{-1}(y)) / dy) (f_X(g^{-1}(y)))$

Functions of Jointly Distributed Random Variables:

$F_S(s) = P(g(X, Y) < s) = \iint_{g(x, y) < s} f_{XY}(x, y)\, dx\, dy$.

Moments: $E((E - E(X))^k)$ is the <u>k^{th} central moment</u>.

The moment generating function is $M_x(t) = E(e^{tx}) = \sum e^{tx} P(X = x) = \int e^{tx} f_x(x)\, dx$.

If Mx(t) exists in an open interval about 0, Mxk(0) is the k^{th} moment of X (f^k is the k^{th} derivative).

If $M_X(t)$ exists in an open interval about 0 then $M_X(t)$ uniquely determines the distribution of X.

If $Y \sim a + bX$ then $M_Y(t) = e^{at} M_X(bt)$.

If X and Y are independent and $S = X + Y$, then $M_S(t) = M_X(t)\, M_Y(t)$.

<u>Statistics</u>

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($\hat{\theta}$) = $E(\hat{\theta}) - \theta$ = expected distance from the mean. A statistic is <u>unbiased</u> if the expectation of the statistic is the true value for the population.
  - $Var(\hat{\theta}) = E((\hat{\theta} - E(\hat{\theta}))^2)$
  - Mean squared error: $MSE = E((\hat{\theta} - \theta)^2) = Var(\hat{\theta}) + bias(\hat{\theta})$.
  - Asymptotically unbiased. $(\lim_{n \to \infty} E(\hat{\theta}) = \theta)$
  - Consistent: $\hat{\theta}$ approaches θ in probability. (ie: $\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$.)

Simple Random Sample (SRS): A sample from a population of size Nis 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 $X_i \sim [\mu_x, \sigma_x^2]$ then X-bar = $\sum X_i/n \sim [\mu_x, (\sigma_x^2/n)(1 - (n-1)/(N-1))]$. (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.) $\lim_{n \to \infty} P((\text{X-bar} - \mu_x)/(\sigma_x/\sqrt{n}) < z) = \Phi(z)$. (The distribution of X-bar approaches $N(\mu_x, \sigma_x^2/n)$.)
- (Law of Large Numbers.) With infinite N, $\lim_{n \to \infty} P(|\text{X-bar} - \mu_x| > t) = 0$ 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 $Cov(X_i, X_j) = -\sigma_x^2 / (N-1)$. [Big population → almost independent; with replacement → independent.]

Estimating Variance: (Not as easy.)

- $(\hat{\sigma})^2 = \sum(X_i - \text{X-bar})^2/n$ [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 $\sum((X_i - \mu)/\sigma)^2 \sim \chi_{n-1}^2$; in other words, $(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$.
  - 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 <u>probability interval</u> for X-bar. Only now X-bar is known and μ_X might not be. Since μ_x is not a random variable, however, this is not a probability interval.
- X-bar $\pm z^*\sigma_{\text{x-bar}}$ is a confidence interval with z from N(0, 1) if $\sigma_{\text{x-bar}}$ is known.
  - $\sigma_{\text{x-bar}} = (\sigma_x/\sqrt{n})(\sqrt{(1-(n-1)/(N-1))})$
  - If σ_x 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 $z^*\sigma_{\text{x-bar}}$ is the <u>margin of error</u>.

Estimating Parameters:

- Method of Moments
  - Let $\mu_k = E(X^k)$; this is the k^{th} moment. We may estimate μ_k by $(\text{X-bar})_k = \sum X_i^k / n$. This is an unbiased estimate.
  - Express parameters as function of $\mu_1, \mu_2, \ldots$ Then use the estimates of μ_k to estimate the parameters. ($\theta_i = f_i(\mu_1, \mu_2, \ldots)$, and $\hat{\theta}_i = f_i(\text{X-bar}_1, \text{X-bar}_2, \ldots)$.)
  - If the f_i are continuous, then $\hat{\theta}$ is a consistent estimator for θ.
- Maximum Likelihood Estimate
  - Suppose $X_1, \ldots, X_n \sim f_x(x \mid \theta)$. (The distribution depends on the values of some parameters.)
  - $L(\theta) = lik(\theta) = \Pi f_x(x_i \mid \theta)$. This is the likelihood of getting this distribution of X_i given a value of the parameter. The θ that maximizes this is the maximum likelihood estimate.
  - The log-likelihood function is: $l(\theta) = \ln(L(\theta)) = \sum \ln(f_x(x_i \mid \theta))$. Maximizing this is equivalent to maximizing $L(\theta)$ and is often easier.
  - The information function is: $I(\theta) = E((\partial (\ln(f_x(x \mid \theta))) / \partial \theta)^2) = -E(\partial^2 (\ln(f_x(x \mid \theta))) / \partial \theta^2)$
  - The variance of $\hat{\theta}$ is $1/nI(\theta)$; this may be estimated with $\hat{\theta}$.

- Maximum likelihood estimates are asymptotically unbiased, consistent and asymptotically normal. If θ is the mle of θ_0 then $\theta \sim N(\theta_0, 1/nI(\theta_0))$ as n grows.
- Cramer-Rao Lower Bound: Let $T(X_1, \ldots, X_n)$ be an unbiased estimator for θ, Then $Var(T(X_1, \ldots, X_n) \geq 1/nI(\theta)$. (This is the best possible variance.)
- For multiple parameters, the maximum likelihood estimates maximize $l(\boldsymbol{\theta})$ for all. (Take all partials and solve.)
- Transformation Invariance: If θ is the mle for θ_0 and g is an invertible function then the mle for $\phi = g(\theta_0)$ is $g(\theta)$.
- *Likelihood Principle.* Two likelihood functions are <u>equivalent</u> 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 θ made from each should be the same.

- Sufficient Statistics
  - A statistic $T(X_1, \ldots, X_n)$ is <u>sufficient</u> for θ if the conditional distribution of $X_1, \ldots, X_n$ given $T(X_1, \ldots, X_n)$ does not depend on θ.
  - Equivalently, $P(X_1 = x_1, \ldots, X_n = x_n$ and $T = t) / P(T = t)$ is not a function of θ.
  - *Factorization Theorem.* $T(X_1, \ldots, X_n)$ is sufficient for θ if and only if $f(x_1,\ldots,x_n \mid \theta) = g(T(x_1, \ldots, x_n), \theta)\, h(x_1, \ldots, x_n)$.
  - 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 $f_X(x|\theta) = \exp(\sum c_j(\boldsymbol{\theta})T(x) + d(\boldsymbol{\theta}) + S(x))\, \delta(x \in A)$, where A does not depend on θ. In this case, $\sum T_j(X_i)$ for each j, are the sufficient statistics for f_x.
  - If $T(\mathbf{X})$ is sufficient for θ then the maximum likelihood estimator of θ is a function of $T(\mathbf{X})$.

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 $T(\mathbf{X})$, some function of the data. If $T(\mathbf{X})$ is in the rejection region, then H_0 is rejected. Otherwise, it is not rejected.
  - The <u>significance level</u> of the test, α, is the probability of $T(\mathbf{X})$ being in the rejection region when H_0 is true (a type I error).
  - The <u>power</u> 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, $T(\mathbf{X}) = L(\theta_0) / L(\theta_a)$.
  - Generalized Likelihood Ratio, when H_0 is $\theta \in \omega_0$ and H_a is $\theta \in \omega_a$:
    - $\Lambda^* = \max(\theta \in \omega_0)\, L(\theta) / \max(\theta \in \omega_a)\, L(\theta)$
    - $\Lambda = \max(\theta \in \omega_0)\, L(\theta) / \max(\theta \in \omega_a \cup \omega_0)\, L(\theta)$

<u>Distributions</u>

(Discrete)

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

$$P(X = x) = \begin{cases} 1/n & \text{if x is a possible outcome} \\ 0 & \text{otherwise} \end{cases}$$

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.

$P(X = x) = p^x (1-p)^{n-x}$ (n *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 ~ 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 ~ Negative Binomial(r, p).

$P(X = x) = (1-p)^{x-r} p^r$ (x-1 *choose* 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 ~ Hypergeometric(N, n, r)

P(X = x) = (r *choose* x) (N-r *choose* n-x) / (N *choose* n)

if X = 0, 1, … , min{n, r}

Poisson: If X is the number of events that occur in a fixed interval, with average rate λ, X ~ Poisson(λ).

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, …

$E(X) = Var(x) = \lambda$

(Continuous)

Continuous Uniform: if X has a uniform probability on [a, b] then X ~ Uniform(a, b)

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

Exponential: X ~ Exp(λ) if

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

Interesting Note: This function is memory-less.

Gamma: X ~ Gamma(α, λ) if

$f_X(x) = (\lambda^\alpha / \Gamma(\alpha)) x^{\alpha-1} e^{-\lambda x}$ where $x \ge 0$

$\Gamma(\alpha) = \int u^{\alpha-1} e^{-u} du$ [from 0 to ∞]

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

$E(X) = \alpha / \lambda$

Normal Distribution: If X ~ N(μ, σ^2) then

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

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

Properties of the normal distribution:

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

Links Between Distributions:

- Bernoulli(p) ~ Binomial(1, p)
- Geometric(p) ~ 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(λ) ~ Gamma(1, λ)
- Standard Normal ~ N(0, 1)
- $\chi^2(n)$ ~ Gamma(n/2, 1/2)
- $\chi^2(n)$ is generated by finding the distribution of $\sum(X_i - \mu_x)^2/\sigma_x^2$ when $X_i \sim N(\mu_x, \sigma_x^2)$.
- If Z ~ 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):

$P(X \le x) = P(Z \le (x - \mu)/\sigma)$