Distributions

Universality of Uniform

When you plug any CRV into its own CDF, you get a Unif$(0,1)$ random variable. When you plug a Unif$(0,1)$ r.v. into an inverse CDF, you get an r.v. with that CDF. Similarly, if $U \sim \text{Unif}(0,1)$ then $F^{-1}(U)$ has CDF F. The key point is that for any continuous random variable X, we can transform it into a Uniform random variable and back by using its CDF.

Binomial Distribution

Let $X \sim \text{Bin}(n,p), Y \sim \text{Bin}(m,p)$ with $X \perp\!\!\!\perp Y$.

- **Redefine success** $n - X \sim \text{Bin}(n, 1-p)$
- **Sum** $X + Y \sim \text{Bin}(n+m, p)$
- **Conditional** $X|(X+Y=r) \sim \text{HGeom}(n,m,r)$ (Fisher exact test)
- **Binomial-Poisson Relationship** $\text{Bin}(n,p)$ is approximately $\text{Pois}(\lambda)$ if p is small.
- **Binomial-Normal Relationship** $\text{Bin}(n,p)$ is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

Hypergeometric Distribution

- **Capture-recapture** A forest has N elk, you capture n of them, tag them, and release them. Then you recapture a new sample of size m. How many tagged elk are now in the new sample? $\text{HGeom}(n, N-n, m)$

Poisson Distribution

Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$.

- **Sum** $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
- **Conditional** $X|(X+Y=n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
- **Chicken-egg** If there are $Z \sim \text{Pois}(\lambda)$ items and we randomly and independently "accept" each item with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.

Normal Distribution

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

CDF and PDF $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ $f(x) = \varphi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}$

Location-Scale Transformation Every time we shift a Normal r.v. (by adding a constant) or rescale a Normal (by multiplying by a constant), we change it to another Normal r.v. For any Normal $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0,1)$ by the following transformation:

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Standard Normal The Standard Normal, $Z \sim \mathcal{N}(0,1)$, has mean 0 and variance 1. Its odd central moments are all 0 as well.

Transformations For constant a, $aX \sim N(a\mu, a^2\sigma^2)$. For $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, $X + Y \sim (\mu + \mu_Y, \sigma^2 + \sigma_Y^2)$

Sum is Normal If X_1 and X_2 are independent and $X_1 + X_2$ is Normal, then X_1 and X_2 must be Normal.

Exponential Distribution

Let us say that X is distributed Expo(λ). We know the following:

CDF

$$F(x) = 1 - e^{-\lambda x}, \text{ for } x \in (0, \infty)$$

Expos as a rescaled Expo(1)

$$Y \sim \text{Expo}(\lambda) \to X = \lambda Y \sim \text{Expo}(1)$$

Memorylessness The Exponential Distribution is the only continuous memoryless distribution. The memoryless property says that for $X \sim \text{Expo}(\lambda)$ and any positive numbers s and t,

$$P(X > s+t | X > s) = P(X > t)$$

Equivalently,

$$X - a | (X > a) \sim \text{Expo}(\lambda)$$

Min of Expos If we have independent $X_i \sim \text{Expo}(\lambda_i)$, then $\min(X_1, \ldots, X_k) \sim \text{Expo}(\lambda_1 + \lambda_2 + \cdots + \lambda_k)$.

Max of Expos If we have i.i.d. $X_i \sim \text{Expo}(\lambda)$, then $\max(X_1, \ldots, X_k)$ has the same distribution as $Y_1 + Y_2 + \cdots + Y_k$, where $Y_j \sim \text{Expo}(j\lambda)$ and the Y_j are independent.

Gamma Distribution

Let us say that X is distributed Gamma(a, λ). We know the following:

Story You sit waiting for shooting stars, where the waiting time for a star is distributed Expo(λ). You want to see n shooting stars before you go home. The total waiting time for the nth shooting star is Gamma(n, λ).

Location-Scale Transformation If $Y \sim \Gamma(a, \lambda)$, then $\lambda Y \sim \Gamma(a, 1)$.

The support is nonnegative and the distribution is right-skewed.

Beta Distribution

Conjugate Prior of the Binomial

$$X|p \sim \text{Bin}(n,p)$$
$$p \sim \text{Beta}(a,b)$$

Then after observing $X = x$, we get the posterior distribution

$$p|(X=x) \sim \text{Beta}(a+x, b+n-x)$$

Bayes' Billiards For any integers k and n with $0 \le k \le n$,

$$\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}$$

Beta-Gamma relationship If $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$, with $X \perp\!\!\!\perp Y$ then

- $\frac{X}{X+Y} \sim \text{Beta}(a,b)$
- $X + Y \perp\!\!\!\perp \frac{X}{X+Y}$

This is known as the **bank–post office result**.

Normalizing Constant $\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

χ^2 (Chi-Square) Distribution

Let us say that X is distributed χ^2_n. We know the following:

Story A Chi-Square(n) is the sum of the squares of n independent standard Normal r.v.s.

Properties and Representations

$$X \text{ is distributed as } Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ for i.i.d. } Z_i \sim \mathcal{N}(0,1)$$
$$X \sim \text{Gamma}(n/2, 1/2)$$

Multinomial Distribution

Let us say that the vector $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_k) \sim \text{Mult}_k(n, \mathbf{p})$ where $\mathbf{p} = (p_1, p_2, \ldots, p_k)$.

Story We have n items, which can fall into any one of the k buckets independently with the probabilities $\mathbf{p} = (p_1, p_2, \ldots, p_k)$.

Note The $X_1, \ldots, X_k$ are dependent.

Joint PMF For $n = n_1 + n_2 + \cdots + n_k$,

$$P(\mathbf{X} = \mathbf{n}) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Marginal PMF, Lumping, and Conditionals Conditioning on some X_j also still gives a Multinomial:

$$X_1, \ldots, X_{k-1} | X_k = n_k \sim \text{Mult}_{k-1}\left(n - n_k, \left(\frac{p_1}{1-p_k}, \ldots, \frac{p_{k-1}}{1-p_k}\right)\right)$$

Variances and Covariances We have $X_i \sim \text{Bin}(n, p_i)$ marginally, so $\text{Var}(X_i) = np_i(1-p_i)$. Also, $\text{Cov}(X_i, X_j) = -np_ip_j$ for $i \ne j$.

Multivariate Normal (MVN) Distribution

A vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ is Multivariate Normal if every linear combination is Normally distributed, i.e., $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ is Normal for any constants $t_1, t_2, \ldots, t_k$. The parameters of the Multivariate Normal are the **mean vector** $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ and the **covariance matrix** where the (i,j) entry is $\text{Cov}(X_i, X_j)$.

Properties The Multivariate Normal has the following properties.

- Any subvector is also MVN.
- If any two elements within an MVN are uncorrelated, then they are independent.
- The joint PDF of a Bivariate Normal (X, Y) with $\mathcal{N}(0,1)$ marginal distributions and correlation $\rho \in (-1, 1)$ is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\tau} \exp\left(-\frac{1}{2\tau^2}(x^2 + y^2 - 2\rho xy)\right),$$

with $\tau = \sqrt{1-\rho^2}$.

Important CDFs

Standard Normal Φ

Exponential(λ) $F(x) = 1 - e^{-\lambda x}$, for $x \in (0, \infty)$

Uniform(0,1) $F(x) = x$, for $x \in (0,1)$

Convolutions of Random Variables

A convolution of n random variables is simply their sum. For the following results, let X and Y be *independent*.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p) \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$. $\text{Bin}(n,p)$ can be thought of as a sum of i.i.d. $\text{Bern}(p)$ r.v.s.
3. $X \sim \text{Gamma}(a_1, \lambda)$, $Y \sim \text{Gamma}(a_2, \lambda)$ $\longrightarrow X + Y \sim \text{Gamma}(a_1 + a_2, \lambda)$. $\text{Gamma}(n, \lambda)$ with n an integer can be thought of as a sum of i.i.d. $\text{Expo}(\lambda)$ r.v.s.
4. $X \sim \text{NBin}(r_1, p)$, $Y \sim \text{NBin}(r_2, p)$ $\longrightarrow X + Y \sim \text{NBin}(r_1 + r_2, p)$. $\text{NBin}(r,p)$ can be thought of as a sum of i.i.d. $\text{Geom}(p)$ r.v.s.
5. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\longrightarrow X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Symmetry

1. If $X \sim \text{Bin}(n, 1/2)$, then $n - X \sim \text{Bin}(n, 1/2)$.
2. If $U \sim \text{Unif}(0,1)$, then $1 - U \sim \text{Unif}(0,1)$.
3. If $Z \sim \mathcal{N}(0,1)$, then $-Z \sim \mathcal{N}(0,1)$. $\varphi(z) = \varphi(-z)$.
4. $\Phi(z) = 1 - \Phi(-z)$

Special Cases of Distributions

1. $\text{Bin}(1,p) \sim \text{Bern}(p)$
2. $\text{Beta}(1,1) \sim \text{Unif}(0,1)$
3. $\text{Gamma}(1, \lambda) \sim \text{Expo}(\lambda)$
4. $\chi^2_n \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$
5. $\text{NBin}(1,p) \sim \text{Geom}(p)$

Moments and MGFs

Moments of Symmetric Distributions

A distribution is symmetric around its mean μ if $X - \mu$ has the same distribution as $\mu - X$. Then for any odd n, the nth central moment is $E(X-\mu)^n$ if it exists. X is symmetric around μ if and only if $f(x) = f(2\mu - x)$ for all x, if f is the PDF of X.

Moment Generating Functions

MGF For any random variable X, the function

$$M_X(t) = E(e^{tX})$$

is the **moment generating function (MGF)** of X, if it exists for all t in some open interval containing 0.

Why is it called the Moment Generating Function? Because the kth derivative of the moment generating function, evaluated at 0, is the kth moment of X.

$$\mu_k = E(X^k) = M_X^{(k)}(0)$$

MGF of linear functions If we have $Y = aX + b$, then

$$M_Y(t) = E(e^{t(aX+b)}) = e^{bt}E(e^{(at)X}) = e^{bt}M_X(at)$$

Uniqueness *If it exists, the MGF uniquely determines the distribution.*

MGF of location-scale transformation If X has MGF $M(t)$, then the MGF of $a + bX$ is $E(e^{t(a+bX)}) = e^{at}M(bt)$

Summing Independent RVs by Multiplying MGFs. If X and Y are independent, then

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X(t) \cdot M_Y(t)$$

The MGF of the sum of two random variables is the product of the MGFs of those two random variables.

Joint Distributions

The **joint CDF** of X and Y is

$$F(x, y) = P(X \leq x, Y \leq y)$$

In the discrete case, X and Y have a **joint PMF**

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In the continuous case, they have a **joint PDF**

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.

Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

$$P(X = x) = \sum_y P(X = x, Y = y)$$

Marginal PDF from joint PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Multivariate LOTUS

LOTUS in more than one dimension is analogous to the 1D LOTUS. For continuous random variables:

$$E(g(X, Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$$

Transformations

One Variable Transformations Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in some function of X. We call this function $Y = g(X)$. Also let $y = g(x)$. If g is differentiable and strictly increasing (or strictly decreasing), then the PDF of Y is

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

Convolutions

Convolution Integral If you want to find the PDF of the sum of two independent CRVs X and Y, you can do the following integral:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)dx$$

Example Let $X, Y \sim \mathcal{N}(0, 1)$ be i.i.d. Then for each fixed t,

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(t-x)^2/2}dx$$

By completing the square and using the fact that a Normal PDF integrates to 1, this works out to $f_{X+Y}(t)$ being the $\mathcal{N}(0, 2)$ PDF.

Poisson Process

Definition We have a **Poisson process** of rate λ arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length t is Pois(λt).
2. Numbers of arrivals in disjoint time intervals are independent.

Count-Time Duality Consider a Poisson process of emails arriving in an inbox at rate λ emails per hour. Let T_n be the time of arrival of the nth email (relative to some starting time 0) and N_t be the number of emails that arrive in $[0, t]$. Let's find the distribution of T_1. The event $T_1 > t$, the event that you have to wait more than t hours to get the first email, is the same as the event $N_t = 0$, which is the event that there are no emails in the first t hours. So

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t} \longrightarrow P(T_1 \leq t) = 1 - e^{-\lambda t}$$

Thus we have $T_1 \sim$ Expo(λ). By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d. Expo(λ), i.e., the differences $T_n - T_{n-1}$ are i.i.d. Expo(λ).

Covariance and Transformations

Covariance and Correlation

Covariance is the analog of variance for two random variables.

$$\text{Cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(X^2) - (E(X))^2 = \text{Var}(X)$$

Correlation is a standardized version of covariance that is always between -1 and 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance and Independence

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X, Y) = 0 \longrightarrow E(XY) = E(X)E(Y)$$

Covariance and Variance The variance of a sum can be found by

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{i<j}\text{Cov}(X_i, X_j)$$

If X and Y are independent then they have covariance 0, so

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

If $X_1, X_2, \ldots, X_n$ are identically distributed and have the same covariance relationships (often by **symmetry**), then

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = n\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$$

Covariance Properties For random variables W, X, Y, Z and constants a, b:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{Cov}(Y, X)\\ \text{Cov}(X + a, Y + b) &= \text{Cov}(X, Y)\\ \text{Cov}(aX, bY) &= ab\text{Cov}(X, Y)\\ \text{Cov}(W + X, Y + Z) &= \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y)\\ &\quad + \text{Cov}(X, Z)\end{aligned}$$

Correlation is location-invariant and scale-invariant For any constants a, b, c, d with a and c nonzero,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

Order Statistics

Definition Let's say you have n i.i.d. r.v.s $X_1, X_2, \ldots, X_n$. If you arrange them from smallest to largest, the ith element in that list is the ith order statistic, denoted $X_{(i)}$. So $X_{(1)}$ is the smallest in the list and $X_{(n)}$ is the largest in the list.

Note that the order statistics are *dependent*, e.g., learning $X_{(4)} = 42$ gives us the information that $X_{(1)}, X_{(2)}, X_{(3)}$ are ≤ 42 and $X_{(5)}, X_{(6)}, \ldots, X_{(n)}$ are ≥ 42.

Distribution Taking n i.i.d. random variables $X_1, X_2, \ldots, X_n$ with CDF $F(x)$ and PDF $f(x)$, the CDF and PDF of $X_{(i)}$ are:

$$F_{X_{(i)}}(x) = P(X_{(i)} \leq x) = \sum_{k=i}^{n}\binom{n}{k}F(x)^k(1 - F(x))^{n-k}$$

$$f_{X_{(i)}}(x) = n\binom{n-1}{i-1}F(x)^{i-1}(1 - F(x))^{n-i}f(x)$$

Uniform Order Statistics The jth order statistic of i.i.d. $U_1, \ldots, U_n \sim$ Unif(0, 1) is $U_{(j)} \sim$ Beta($j, n - j + 1$).

Conditional Expectation

Conditioning on an Event We can find $E(Y|A)$, the expected value of Y given that event A occurred. A very important case is when A is the event $X = x$. Note that $E(Y|A)$ is a *number*.

$$E(Y|A) = \int_{-\infty}^{\infty} yf(y|A)dy$$

Conditioning on a Random Variable We can also find $E(Y|X)$, the expected value of Y given the random variable X.

Properties of Conditional Expectation

1. $E(Y|X) = E(Y)$ if $X \perp\!\!\!\perp Y$
2. $E(h(X)W|X) = h(X)E(W|X)$ (**taking out what's known**) In particular, $E(h(X)|X) = h(X)$.
3. $E(E(Y|X)) = E(Y)$ (**Adam's Law**, a.k.a. Law of Total Expectation)

Adam's Law with Extra Conditioning

$$E(E(Y|X, Z)|Z) = E(Y|Z)$$

Eve's Law (a.k.a. Law of Total Variance)

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

MVN, LLN, CLT

Sample mean

Let $X_1, X_2, X_3 \dots$ be i.i.d. with mean μ. The **sample mean** is

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

. Then, $E(\bar{X}_n) = \mu$ and $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

Law of Large Numbers (LLN)

The **Law of Large Numbers** states that as $n \to \infty$, $\bar{X}_n \to \mu$ with probability 1. For example, in flips of a coin with probability p of Heads, let X_j be the indicator of the jth flip being Heads. Then LLN says the proportion of Heads converges to p (with probability 1).

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable $Y = X_1 + X_2 + \cdots + X_n$ that is a sum of n i.i.d. random variables X_i. Let $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma_Y^2$. The CLT says

$$Y \dot{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$$

If the X_i are i.i.d. with mean μ_X and variance σ_X^2, then $\mu_Y = n\mu_X$ and $\sigma_Y^2 = n\sigma_X^2$. For the sample mean $\bar{X}_n$, the CLT says

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \dot{\sim} \mathcal{N}(\mu_X, \sigma_X^2/n)$$

Asymptotic Distributions using CLT

We use $\xrightarrow{D}$ to denote *converges in distribution to* as $n \to \infty$.

$$\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{D} \mathcal{N}(0, 1)$$

$$\sqrt{n}(\bar{X}_n - \mu_X) \xrightarrow{D} \mathcal{N}(0, \sigma_X^2)$$

Markov Chains

Definition

A Markov chain must satisfy the **Markov property**, which says that if you want to predict where the chain will be at a future time, if we know the present state then the entire past history is irrelevant. *Given the present, the past and future are conditionally independent.* In symbols,

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

Transition Matrix

Let the state space be $\{1, 2, \dots, M\}$. The transition matrix Q is the $M \times M$ matrix where element q_{ij} is the probability that the chain goes from state i to state j in one step:

$$q_{ij} = P(X_{n+1} = j | X_n = i)$$

To find the probability that the chain goes from state i to state j in exactly m steps, take the (i, j) element of Q^m.

$$q_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

If X_0 is distributed according to the row vector PMF $\mathbf{p}$, i.e., $p_j = P(X_0 = j)$, then the PMF of X_n is $\mathbf{p}Q^n$. The number of free parameters in this system depends on how many free parameters are in the transition matrix.

Chain Properties

A chain is **irreducible** if you can get from anywhere to anywhere. If a chain (on a finite state space) is irreducible, then all of its states are recurrent. A chain is **periodic** if any of its states are periodic, and is **aperiodic** if none of its states are periodic. In an irreducible chain, all states have the same period.

A chain is **reversible** with respect to $\mathbf{s}$ if $s_i q_{ij} = s_j q_{ji}$ for all i, j. Examples of reversible chains include any chain with $q_{ij} = q_{ji}$, with $\mathbf{s} = (\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M})$, and random walk on an undirected network.

Stationary Distribution

Let us say that the vector $\mathbf{s} = (s_1, s_2, \dots, s_M)$ be a PMF (written as a row vector). We will call $\mathbf{s}$ the **stationary distribution** for the chain if $\mathbf{s}Q = \mathbf{s}$.

For irreducible, aperiodic chains, the stationary distribution exists, is unique, and s_i is the long-run probability of a chain being at state i. The expected number of steps to return to i starting from i is $1/s_i$.

To find the stationary distribution, you can solve the matrix equation $(Q' - I)\mathbf{s}' = 0$. The stationary distribution is uniform if the columns of Q sum to 1. This is true for symmetric matrices

Reversibility Condition Implies Stationarity If you have a PMF $\mathbf{s}$ and a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all states i, j implies that $\mathbf{s}$ is stationary.

Columns Summing to One Implies Stationarity If each column of the transition matrix Q sums to 1, then the uniform distribution over all the states, $(1/M, 1/M, \dots, 1/M)$ is a stationary distribution. One example of this is a symmetric transition matrix.

Random Walk on an Undirected Network

The stationary distribution of a random walk chain is proportional to the **degree sequence** (this is the sequence of degrees, where the degree of a node is how many edges are attached to it). For example, the stationary distribution of random walk on the network shown above is proportional to $(3, 3, 2, 4, 2)$, so it's $(\frac{3}{14}, \frac{3}{14}, \frac{3}{14}, \frac{4}{14}, \frac{2}{14})$.

Inequalities

1. **Cauchy-Schwarz** $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$
2. **Markov** $P(X \geq a) \leq \frac{E|X|}{a}$ for $a > 0$
3. **Chebyshev** $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$ for $E(X) = \mu, \text{Var}(X) = \sigma^2$. Useful for proving convergence in probability; to prove consistency of estimator we just need to show that its variance goes to 0.
4. **Jensen** $E(g(X)) \geq g(E(X))$ for g convex; reverse if g is concave
5. **Chernoff's** For any r.v. X with finite mean μ and constant $a, t > 0$, $P(X \geq a) \leq \frac{E(e^t X)}{e^{ta}}$

Formulas

1. **Geometric Series**
   $1 + r + r^2 + \cdots + r^{n-1} = \sum_{k=0}^{n-1} r^k = \frac{1-r^n}{1-r}$
   $1 + r + r^2 + \cdots = \frac{1}{1-r}$ if $|r| < 1$
2. $\mathbf{e^x} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$
3. **Binomial Theorem**
   $(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$
4. **Gamma and Beta Integrals**
   $\int_0^\infty x^{t-1} e^{-x}\, dx = \Gamma(t)$ $\quad$ $\int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
   Also, $\Gamma(a + 1) = a\Gamma(a)$, and $\Gamma(n) = (n - 1)!$ if n is a positive integer.
5. $\sum_n (X_j - \bar{X})(Y_j - \bar{Y}) = \sum_n X_j Y_j - n\bar{X}\bar{Y}$
6. $nE(\bar{X}\bar{Y}) = \frac{1}{n}\sum_{i,j} E(X_i Y_i) = \frac{1}{n}(nE(X_1 Y_1) + n(n-1)E(X_1 Y_2))$

STAT 111 Stuff

Summary Statistics

Medians and Quantiles Let X have CDF F. Then X has median m if $F(m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. For X continuous, m satisfies $F(m) = 1/2$. In general, the ath quantile of X is $\min\{x : F(x) \geq a\}$; the median is the case $a = 1/2$.

Standard Error $SE(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$

Likelihood

$$L(\theta; \mathbf{y}) = f_{\mathbf{Y}}(y|\theta)$$

It is regarded as a function of θ, with y treated as fixed.

Frequentist Interpretation θ is regarded as fixed but unknown, and it does not have a distribution.

Bayesian Interpretation We have a prior density $g(\theta)$ for θ, then:

$$L(\theta) = g(\theta|\mathbf{y} = \frac{g(\theta)f(y|\theta)}{f(\mathbf{y})} \propto g(\theta)f(y|\theta) = L(\theta)g(\theta)$$

So the posterior is proportional to likelihood times prior.

Equivalence Two likelihood functions are viewed as equivalent if one is a positive constant times the other. In fact, the "constant" can even be function of the data (it just can't depend on the parameter).

Invariance Let $\psi = g(\theta)$ be a reparametrization, where g is a one-to-one function. Then $L(\psi; \mathbf{y} = L(\theta; \mathbf{y})$.

Estimands, Estimators, & Estimates

Estimand An estimand is an object that we wish to learn about from data.

Estimator An estimator $\hat{\theta} = T(\mathbf{Y})$ is a statistic with the intention of estimating an estimand θ.

Estimate An estimate is a realization of an estimator. If $T(Y)$ is an estimator of some estimand θ, then $T(y)$ is an estimate of θ.

Method of Moments

Set the expectation of the sample moments equal to the actual sample moments. Compute the expectation in terms of the unknown parameter(s) and rearrange to get the estimator. Write as many equations as you have parameters, one equation per moment. For example, suppose you have unknown parameters θ and λ.

$$\text{1st moment } E(\frac{1}{n}\sum X_i) = f(\theta, \lambda)$$

$$\text{2nd moment } E(\frac{1}{n}\sum X_i^2) = f(\theta, \lambda)$$

You can then solve this system of equations for $\hat{\theta}$ and $\hat{lambda}$.

Maximum Likelihood Estimation

The **maximum likelihood estimate** of θ is the value $\hat{\theta}$ that maximizes the likelihood function $L(\theta; y)$.

Regularity conditions Support must not depend on the value of the estimand. The estimate θ^* must not be on the boundary. You must be able to Differentiate under the Integral Sign, i.e. $\frac{d}{d\theta}\int g(y) f_\theta(y) dy = \int \frac{d}{d\theta} g(y) f_\theta(y) dy$. The estimand must be of a fixed dimension.

Invariance If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $\hat{\theta}$.

Consistency The MLE $\hat{\theta}$ is consistent, which means that it converges in probability to the true θ.

Asymptotically Normal The MLE is asymptotically Normal (so its distribution is approximately Normal if the sample size is large).

Asymptotically unbiased The MLE is asymptotically unbiased (the bias approaches 0 as the sample size grows).

Asymptotically efficient The MLE is asymptotically efficient (no other asymptotically unbiased estimator will have a lower standard error asymptotically).

Bias, Variance and Loss Functions

Bias The bias of an estimator $\hat{\theta}$ for θ is $E(\hat{\theta}) - \theta$.

Loss function A loss function is a function $L(\theta, \hat{\theta})$, interpreted as the loss associated with using the estimate $\hat{\theta}$ when the true parameter value is θ. We require that $L(\hat{\theta}, \theta) \geq 0$ and $L(\theta, \theta) = 0$.

Mean squared error The mean squared error is $E(\hat{\theta} - \theta)^2$.

Bias-variance decomposition We know $\text{MSE}_\theta = \text{Var}_\theta(\hat{\theta}) + (\text{Bias}_\theta(\hat{\theta}))^2$. This illustrates the *bias-variance tradeoff*.

Kernel Density Estimation

Suppose the estimand is the density of Y_1 at a particular point y_1: $\theta = f_{Y_1}(y_1)$. The kernel density estimator is:

$$\hat{f}_n(y) = \frac{1}{n}\frac{1}{h}K(\frac{Y_i - y}{h})$$

where $h > 0$ is called the *bandwidth* and K is called the *kernel function*. The kernel function must be a PDF (nonnegative and summing to 1). For instance, the Gaussian kernel takes K to be a Normal PDF centered at 0.

Asymptotics and Information

Consistency

An estimator $\hat{\theta}$ is **consistent** for the estimand θ if $\hat{\theta}$ converges in probability to the true θ as the sample size $n \to \infty$, i.e. for every $\epsilon > 0$, we have $\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0$.

Sufficient condition If $\text{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is consistent. In particular, if $\text{Bias}(\hat{\theta}) \to 0$ and $\text{Var}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is consistent.

Kullback-Leibler Divergence

The **Kullback-Leibler divergence** is a way to compare two distributions; it measures the impact on expected log-likelihood if we use an approximate distribution as a proxy for the true distribution. It is defined to be:

$$K(\theta^*, \theta) = E\left(\log \frac{L(\theta^*; \mathbf{Y})}{L(\theta; \mathbf{Y})}\right) = E(\log(L(\theta^*; \mathbf{Y}) - E(\log(L(\theta; \mathbf{Y})$$

where the expectation is computed under the distribution $\mathbf{Y} \sim F_{\mathbf{Y}}(\mathbf{y}|\theta^*)$.

Nonnegative For any θ, we have $K(\theta^*, \theta) \geq 0$. The inequality is strict unless $F_{\mathbf{Y}}(\mathbf{y}|\theta^*)$ and $F_{\mathbf{Y}}(\mathbf{y}|\theta)$ are the same distribution. In particular, θ^* maximizes the expected log-likelihood. the MLE $\hat{\theta}$ is where the *observed* log-likelihood function has its peak, while θ^* is where the *expected* log-likelihood function has its peak, thus lending support for using MLEs. Note that graders may require you to check the second derivative $l''(\theta) < 0$ to confirm you have found a maximum and not a minimum.

Score function and Fisher information

Score function The score function is $s = \frac{\partial l(\theta;y)}{\partial \theta}$.
We have that:

$$E(s(\theta^*; \mathbf{Y})) = 0$$
$$\text{Var}(s(\theta^*; \mathbf{Y}) = -E(s'(\theta^*; \mathbf{Y}))$$

Fisher information The Fisher information for a parameter θ is:

$$I(\theta) = \text{Var}_\theta s(\theta; \mathbf{Y})$$

where the subscript of θ indicates that we compute the variance under the assumption that the true parameter value is θ. We will sometimes write $I_n(\theta)$ for the Fisher information when the sample size is n.

Fisher information of function of r.v. Let $\tau = g(\theta)$, where g is a differentiable function with $g'(\theta) \neq 0$. Then:

$$I(\tau) = \frac{I(\theta)}{(g'(\theta))^2}$$

Cramer-Rao Lower Bound

Let $\hat{\theta}$ be an unbiased estimator of θ. Under regularity conditions,

$$Var(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta^*)}$$

. Since bias is 0, variance = MSE of the estimator. For θ which may be biased, this becomes

$$Var(\hat{\theta}) \geq \frac{g'(\theta^*))^2}{\mathcal{I}(\theta^*)}$$

where $E(\hat{\theta}) = g(\theta^*)$.

Asymptotic distribution of the MLE

For large sample size, it is *approximately* true that the MLE is Normal, unbiased, and achieves the CRLB.
Under regularity conditions, the asymptotic distribution of $\hat{\theta}$ is given by:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I_1(\theta^*)}\right)$$

as the sample size $n \to \infty$. As an *approximation*, the result says that for large n,

$$\hat{\theta} \sim \mathcal{N}\left(\theta^*, \frac{1}{nI_1(\theta^*)}\right)$$

Delta Method

The *delta method* says that if:

$$\sqrt{n}(\hat{\theta} - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

and g is a differentiable function, then:

$$\sqrt{n}(g(\hat{\theta}) - g(\mu)) \xrightarrow{D} \mathcal{N}(0, (g'(\mu))^2\sigma^2)$$

Interval Estimation

Use a pivot to write a function of the estimator of interest that has a distribution whose parameters are known (ex. changing $X \sim \text{Expo}(\lambda)$ to $\lambda X \sim \text{Expo}(1)$. This known distribution also has a known CDF whose function you can use to construct the confidence interval.

Sufficient Statistics

Definition of Sufficient Statistics

Let $\mathbf{Y} = (Y_1, ..Y_n)$ be a sample from model $F_y(y|\theta)$. A statistic $T(Y)$ is a sufficient statistic for θ if conditional distribution of $Y|T$ does not depend on θ.

Factorization Criterion

T(Y) is a sufficient statistic if and only if we can factor

$$f_y(y|\theta) = g(T(y), \theta)h(y)$$

where $f_y(y|\theta)$ is the PMF/PDF of Y.

Rao-Blackwell

Let T be a sufficient statistic and $\hat{\theta}$ be any estimator for θ. Then the MSE of the Rao-Blackwellized estimator $\hat{\theta}_R B = E(\theta|T)$ does not exceed the MSE of the original estimator. This can be proven with the bias-variance decomposition and Adam and Eve's Laws.

Natural Exponential Family

An r.v. follows NEF if its PDF is in the form

$$f_y(y|\theta) = e^{\theta y - \Psi(\theta)}h(y)$$

Where θ is the natural parameter. Note that θ doesn't necessarily have to be a parameter of interest, e.g. it could be $-\mu$ instead of μ for a normal distribution.

Properties

If Y is in NEF form (e.g. Normal (σ^2 known), Poisson, Binomial (n fixed), Negative Binomial (r fixed), $\Gamma(a, \lambda)$ (a known), then we have the following facts.

1. $E(Y) = \Psi'(\theta), \text{Var}(\theta) = \Psi''(\theta)$, MGF $M_y(t) = E(e^{tY}) = e^{\Psi(\theta+t)-\Psi(\theta)}$.
2. $\bar{Y}$ is a sufficient statistic for θ.
3. MLE for mean paramter $\mu = E(Y)$ is $\mu = \bar{Y}$.
4. Fisher Information $I_1(\theta) = \Psi''(\theta)$.

MLEs & Fisher Informations

- Bernoulli: $\hat{p} = \bar{Y}$. $I_1(p) = \frac{1}{pq}$
- Binomial: $\hat{p} = \frac{y}{n}$. $I_n(p) = \frac{n}{pq}$
- Geometric: $\hat{p} = n/\sum y_i$. $I_n(p) = n(\frac{1}{p^2} + \frac{1}{pq})$
- Negative binomial: $\hat{p} = \frac{r}{Y+r}$. $I_1(p) = \frac{r}{q^2 p}$
- Poisson: $\hat{\lambda} = \frac{1}{n}\sum y_i$. $I_1(\lambda) = \frac{1}{\lambda}$
- Exponential: $\hat{\lambda} = n/\sum y_i = 1/\bar{Y}$. $I_1(\lambda) = \lambda^{-2}$
- Normal: $\hat{\mu} = \bar{Y}$. $\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{Y})^2$
- Weibull: $\hat{\lambda}|\gamma = \frac{1}{n}\sum y_i^\gamma$

Student-t Distribution

Let $T = \frac{E}{\sqrt{V/n}}$ where $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_n$ (remember, special case of Gamma) where $Z \perp\!\!\!\perp V$, then $T \sim \text{t}_n$ (where n is called the # of degrees of freedom). If $n = 1$, then $T \sim$ Cauchy. Can find mean, variance, in your head by remembering it's a sum of standard Normals.

Also: For $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, we can find using JFI that $\sum_{j=1}^n (Z_j - \bar{Z_n})^2 \sim \chi^2_{n-1}$.

Linear Regression

Regression function $r(x) = E(Y|X = x) = \int y f(y|x)\, dy$.
Assume data is $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$, i.i.d. $F_{Y,X}$

Simple Linear Regression Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
where $E(\epsilon_i|X_i = x) = 0$ and $\text{Var}(\epsilon_i|X_i = x) = \sigma^2$
Predicted or fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$

Residual Sums of Squares $\sum_{i=1}^n \hat{\epsilon}_i^2$

Least Squares Estimators MLEs are assuming that errors are normally distributed

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \text{ (MLE)}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ (MLE)}$$
$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2 \text{(unbiased, not MLE)}$$

Asymptotic Distribution $\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\sigma_X^2}\right)$

Student-t Distribution Let $\hat{\sigma}^2$ as above, and let $s_{xx}^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. Then

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/s_{xx}} \sim t_{n-2}$$

Prediction We see a new $X = x_0$ and want to predict new Y. Predict Y with estimators, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

$$E(\hat{Y} - Y|\mathbf{X} = \mathbf{x}) = 0$$
$$\text{Var}(\hat{Y} - Y|\mathbf{X} = \mathbf{x}) = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x x^2} + 1\right)$$
$$\frac{\hat{Y} - Y}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x x^2} + 1}} \sim t_{n-2}$$

The $1 - \alpha$ confidence interval for Y is

$$\hat{Y} \pm t_{n-2}^{-1}(\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x x^2} + 1}$$

Slutsky's and Continuous Mapping Theorem

Slustky's Theorem If $X_1, X_2 \dots$ and $Y_1, Y_2, \dots$ are sequences of random variables such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $X_n Y_n \xrightarrow{d} cX$
3. if $c \neq 0$, then $X_n / Y - n \xrightarrow{d} X/c$

Continuous Mapping Theorem If $X_1, X_2, \dots$ are sequences of random variables and g is a continuous function, then

1. if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.
2. if $X_n \xrightarrow{p} c$, then $g(X_n) \xrightarrow{p} g(c)$.

Hypothesis Testing

Null and Alternative Hypotheses

Hypothesis Testing Framework Partition parameter space Θ into two disjoint sets, $\Theta = \Theta_0 \cup \Theta_1$

Null Hypothesis $H_0 : \theta \in \Theta_0$. Simple if it consists of a single point. Composite if not simple.

Alternative Hypothesis $H_1 : \theta \in \Theta_1$.

Rejection Region Set R of possible values $\mathbf{y}$ for the data such that we reject the null hypothesis $\mathbf{y} \in R$. Usually of the form $R = \{\mathbf{y} : t(\mathbf{y}) > c\}$ or $R = \{\mathbf{y} : t(\mathbf{y}) > c_U \text{ or } t(\mathbf{y}) < c_L\}$. Often, c is some quantile, $F^{-1}(\alpha)$.

Type I Error $\theta \in \Theta_0$ but $\mathbf{y} \in R$ (incorrectly reject null).

Type II Error $\theta \in \Theta_1$ but $\mathbf{y} \notin R$ (incorrectly fail to reject).

One-sided test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

Two-sided test $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.

Size and Power

Power function $\beta(\theta) = P(\mathbf{Y} \in R|\theta)$. How likely we are to reject for a given true parameter value. Typically, power of a test refers to $\beta(\theta)$ for $\theta \in \Theta_1$.

Size Size or level of a test is the maximum possible Type I error probability. $\alpha = \max_{\theta \in \Theta_0} \beta(\theta)$.

z-test

!!!!!!!FINISH THIS!!!!!!

t-test

Let

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$T(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

$$\implies T(\mathbf{Y}) \sim t_{n-1}$$

p-values

A **p-value** is the smallest α (Type I error rate) at which we could have rejected H_0, so

$$p = \inf \alpha : y \in R_\alpha$$

Often, our rejection region is of the form $R = \{\mathbf{y} : T(\mathbf{y}) > F^{-1}(\alpha)\}$. In this case, we can calculate $p = F(t(\mathbf{y}))$.
Theorem: Let T be a continuous test statistic, and suppose we are using a hypothesis test that rejects $H_0 : \theta = \theta_0$ when T is large. As a random variable, the p-value is Unif$(0, 1)$ under H_0.

Misc Tests

Wald Test Under some regularity conditions, $\hat{\theta} \dot{\sim} \mathcal{N}(\theta_0, I^{-1}(\theta_0))$. So, can use the test statistic,

$$t(\mathbf{Y}) = \sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, 1)$$

Score test Under some regularity conditions, $s(\mathbf{y}, \theta_0)$ is asymptotically Normal.

$$\frac{s(\mathbf{y}; \theta_0)}{\sqrt{\mathcal{I}(\theta_0)}} \sim \mathcal{N}(0, 1)$$

Likelihood Ratio Test (LRT) If testing a simple null vs simple alternative, then Neyman-Pearson lemma says most powerful test will be based on likelihood ratio

$$LR = \frac{L(\theta_1, \mathbf{Y})}{L(\theta_0, \mathbf{Y})}$$

The more general LRT is given by

$$LR = \frac{L(\hat{\theta}; \mathbf{Y})}{L(\theta_0; \mathbf{Y})}$$

We also have the folllowing theorem under some mild regularity theorems,

$$\Lambda(\mathbf{Y}) = 2\log\left(\frac{L(\hat{\theta}, \mathbf{Y})}{L(\theta_0; \mathbf{Y})}\right) \xrightarrow{D} \chi^2_1$$

Causal Inference

Potential Outcomes Treat $X \in \{0, 1\}$ is treatment assignment. We observe potential outcomes

$$Y = \begin{cases} Y(1) & \text{if } X = 1 \\ Y(0) & \text{if } X = 0 \end{cases}$$

Average Causal Effect $\theta = E[Y(1) - Y(0)]$

Association $\alpha = E[Y(1) \mid X = 1] - E[Y(0) \mid X = 0]$
In general, $\alpha \neq \theta$

Unbiased Estimator $\hat{\theta} = \bar{Y}_1 - \bar{Y}_0 = \hat{\beta}_1$

Sampling

Simple Random Sampling (SRS)

Definition An SRS of size n from a population of size N is a random sample of size n, chosen without replacement, such that all possible samples are equally likely. Y_j is jth individual sampled, y_i is person with label i in population.

Sample Mean Sample mean is unbiased

$$E(\bar{Y}) = \mu$$

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Covariance

$$\text{Cov}(Y_i, Y_j) = \frac{-\sigma^2}{N-1}$$

Stratified Random Sampling

Definition Suppose that the population is partitioned into L subpopulations called *strata*. Let N_ℓ be the size of stratum ℓ, so $N_1 + \dots + N_L = N$. Assume $N_\ell \geq 2 \forall \ell$. Within each stratum ℓ, an SRS of size n_ℓ is drawn, for soome predetermined n_ℓ. Let $y_{i,\ell}$ be the y values within the stratum ℓ, let μ_ℓ be the population mean within stratum ℓ and let σ^2_ℓ be the population variance in stratum ℓ,

$$\mu_\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} y_{i,\ell} \text{ and } \sigma^2_\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell}(y_{i,\ell} - \mu_\ell)^2$$

Estimator The stratified sampling estimator $\bar{Y}_{\text{strat}}$ is

$$\bar{Y}_{\text{strat}} = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \cdot \bar{Y}_\ell$$

This estimator is unbiased. The variance is:

$$\text{Var}\left(\bar{Y}_{\text{strat}}\right) = \sum_{\ell=1}^{L}\left(\frac{N_\ell}{N}\right)^2 \cdot \frac{\sigma^2_\ell}{n_\ell} \cdot \frac{N_\ell - n_\ell}{N_\ell - 1}$$

Optimal Allocation In terms of minimizing MSE, the optimal allocation of sampling is

$$n_\ell = \frac{n N_\ell \tilde{\sigma}_\ell}{\sum_{k=1}^{L} N_k \tilde{\sigma}_k}$$

$$\implies n_\ell \propto N_\ell \tilde{\sigma}_\ell$$

Where $\tilde{\sigma}_\ell$ is the standard deviation in stratum ℓ except defined with $N_\ell - 1$ in the denominator rather than N_ℓ. If there is a cost per sample c_ℓ in each stratum and a total budget, then $n_\ell \propto \frac{N_\ell \tilde{\sigma}_\ell}{\sqrt{c_\ell}}$.

Horvitz-Thompson Estimator

Let S be the set of distinct ID numbers members of a random sample from the population. Let $\pi_i = P(i \in S)$ be the probability that individual i is included in the sample. Assume that π_i are known in advance and that $\pi_i > 0 \forall i$, then the *Horvitz-Thomspon* estimator

$$\tau_{\text{HT}} \sum_{i \in S} \frac{y_i}{\pi_i}$$

is **unbiased** for the population total $\tau = y_1 + \dots + y_N$.
If N is known, then

$$\hat{\mu}_{\text{HT}} = \frac{\hat{\tau}_{\text{HT}}}{N}$$

is an unbiased estimator for the population mean μ. However, while the the Horvitz-Thompson estimator is always unbiased, **it can have very large variance, so is not always best for MSE**.

Simulation Tests

Permutation Tests

Procedure Let $X_1, \dots, X_m \overset{\text{i.i.d.}}{\sim} F_X$ and $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} F_Y$ be independent samples. Want to test $H_0 : F_X = F_Y$ vs. $H_1 : F_X \neq F_Y$.

1. Let T be a test statistic.
2. Compute observed value t_0 from data.
3. Compute T for each permutation of $X_1, \dots, X_m, Y_1, \dots, Y_n$. Call these $t_1, \dots, t_{(m+n)!}$.
4. The p-value based on choosing a random permutation is

$$P(T \geq t_0) = \frac{1}{(m+n)!}\sum_{j=1}^{(m+n)!} I(t_j \geq t_0)$$

There are often a lot of permutations, so can sample permutations randomly instead.

Strengths Much simpler in general.

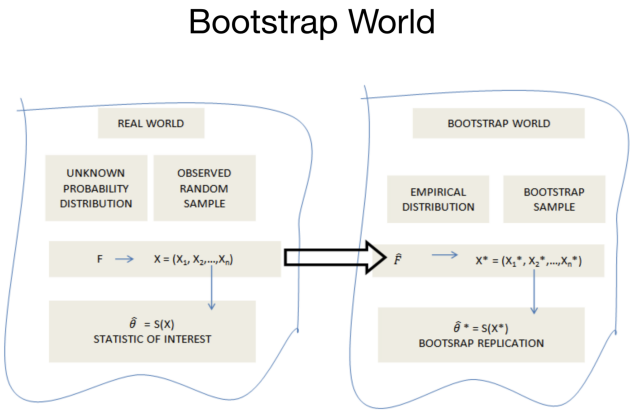
- Can use *any* test statistic you want
- No complicated math, p-value is just a proportion
- No parametric assumptioins
- No asymptotics needed

Limitations Only compares distributions, not parameters.

- Strong, inflexible null hypothesis (equal *distribution*)
- Assumes exchangeability within each group
- General parametric vs. nonparametric considerations

Bootstrap

Resample from sample *with replacement*, use empirical distribution to approximate true distribution.



Sharp Null and Randomization Tests

Sharp Null n units are randomized for one treatment ($X = 1$) or another treatment ($X = 0$). Sharp null hypothesis is

$$H_0 : y_i(1) = y_i(0) \forall i = 1, \dots, n$$
$$H_1 : y_i(1) \neq y_i(0) \text{ for some } i \in \{1, \dots, n\}$$

In words, the sharp **null** hypothesis says that *the treatment has no effect on any individual's outcome* and the **alternative** is that *the treatment has an effect for at least one individual.*

Randomization Tests Compute all possible ways individuals could have been randomized according to randomization protocol. For each of these ways, compute some test statistic. The p-value is the proportion of these test statistics that equal or exceed the observed test statistic.

Bayesian Statistics

Basics

Set up a probability model for quantities of interest in a problem, both known and unknown. Then condition on the observed data, to obtainthe posterior distribution. Two simple rules

1. Always obey the laws of probability
2. All uncertainty is to be modeled using probability

Often controversial is the choice of the prior distribution from which we are shifting.

Theorem: Consider a parametric model $f(\mathbf{y}|\theta)$ for data $\mathbf{y}$, and let $\pi(\theta)$ be the prior density on the parameter θ. Let $L(\theta \mid \mathbf{y})$ be the likelihood function. Then the posterior density of θ is proportional to the likelihood times the prior.

$$\pi(\theta|\mathbf{y}) \propto L(\theta|\mathbf{y})\pi(\theta)$$

Point Estimation

Mean

$$\text{Prior: } E(\theta) = \int_{-\infty}^{\infty} \theta\pi(\theta)\, d\theta$$
$$\text{Posterior: } E(\theta|\mathbf{y}) = \int_{-\infty}^{\infty} \theta\pi(\theta|\mathbf{y})\, d\theta$$

Median Value m such that

$$\text{Prior: } P(\theta \leq m) \int_{-\infty}^{m} \pi(\theta)\, d\theta = \frac{1}{2}$$
$$\text{Posterior: } (\theta \leq m|\mathbf{y}) \int_{-\infty}^{m} \pi(\theta|\mathbf{y})\, d\theta = \frac{1}{2}$$

Mode The prior mode is the value of θ that maximizes $\pi(\theta)$, if this value exists and is unique. The posterior mode is the value of θ that maximizes $\pi(\theta|\mathbf{y})$

Squared Error Loss $C(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Expected squared error loss minimized by the **posterior mean** $E(\theta|\mathbf{y})$.

Absolute Error Loss $C(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. Expected absolute error loss minimized by the **posterior median**.

Credible Intervals

Definition Let $0 < \alpha < 1$. A $1 - \alpha$ credible interval or posterior probabiltiy interval for a parameter θ is an interval estimator $[a(\mathbf{Y}), b(\mathbf{Y})]$ such that

$$P(a(\mathbf{y}) \leq \theta \leq b(\mathbf{y})|\mathbf{y}) = 1 - \alpha$$

Often simpler than confidence intervals, because can just look at CDF of posterior distribution.

Conjugate Priors

Definition A family of priors is *conjugate* for a particular model if choosing a prior in the family always results in a posterior that is in the same family.

Beta-Binomial Conjugacy

Suppose we have

$$Y|p \sim \text{Bin}(n, p)$$

with prior

$$p \sim \text{Beta}(a, b)$$

Then the posterior is still Beta,

$$p|(Y = y) \sim \text{Beta}(a + y, b + n - y)$$

Poisson-Gamma Conjugacy

Suppose we have

$$Y|\lambda \sim \text{Pois}(\lambda t)$$
$$\lambda \sim \text{Gamma}(r_0, b_0)$$

Then, the posterior distribution is

$$\lambda|(Y = y) \sim \text{Gamma}(r_0 + y, b_0 + t)$$

The marginal distribution of Y is given by

$$Y \sim \text{NBin}\left(r_0, \frac{b_0}{b_0 + t}\right)$$

Normal-Normal Conjugacy

The conjugate prior of mean of normal with variance know is also normal, so suppose we have

$$Y|\mu \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

If we define

$$B = \frac{\sigma^2}{\sigma^2 + \tau_0^2}$$

Then the posterior distribution is

$$\mu|(Y = y) \sim \mathcal{N}((1 - B)y + B\mu_0, B\tau_0^2)$$

and the marignal distribution of Y is

$$Y \sim \mathcal{N}(\mu_0, \sigma^2 + \tau_0^2)$$

The posterior mean is a weighted average of the sample mean $\bar{y}$ and the prior mean μ_0. We call B the *shrinkage factor*.

NEF Conjugacy

Let $Y_1, \dots, Y_n$ follow the NEF

$$f(y|\theta) = e^{\theta y - \psi(\theta)} h(y)$$

Assume that $Y_1, \dots, Y_n$ are conditionally independent given θ, so the likelihood functon is

$$L(\theta|y) = e^{n(\theta\bar{y} - \psi(\theta))}$$

Then a conjugate prior on θ is

$$\phi(\theta) \propto e^{r_0(\theta\mu_0 - \psi(\theta))}$$

Furthermore, the posterior mean of the mean parameter

$$\mu = E(Y_1|\theta) = \psi'(\theta)$$

is a weighted average of the sample mean and the prior mean

$$E(\mu|y) = (1 - B)\bar{y} + B\mu_0$$

where

$$B = \frac{r_0}{r_0 + n}$$

Stein's Paradox

Risk function Given a cost function $C(\theta, \hat{\theta})$, the risk function of an estimator $\hat{\theta}$ is its expected loss

$$R(\theta) = E(C(\theta, \hat{\theta})|\theta)$$

Admissibility An estimator $\hat{\theta}$ is *inadmissible* if there exists another estimator whose risk fucntion is less than or equal to that of $\hat{\theta}$ for all possible θ, with strict inequality for at least one possible value of θ. An estimator is *admissible* if it is not inadmissible.

Inadmissibility of MLE Let $Y_i \sim \mathcal{N}(\mu_i, V)$ for $i = 1, \dots, k$ be independent, where $k \geq 3$ and μ_i unknown and V known. Let the estimand be $\boldsymbol{\mu} = (\mu_i, \dots, \mu_k)$ and the loss function be total squared error loss $C(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{k} (\mu_i - \mu)^2$. Then the MLE, which is $\mathbf{Y} = (Y_1, \dots, Y_k)$ is inadmissible.

James-Stein estimator Let $S = \sum_{i=1}^{k} Y_i^2$. Then we have the James-Stein estimator:

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(k-2)V}{S}\right) Y_j$$

where $\hat{\boldsymbol{\mu}}_{\text{JS}}$ has strictly lower risk that $\mathbf{Y}, \forall \boldsymbol{\mu} \in \mathbb{R}^k$. Specifically, risk functions are

$$R(\boldsymbol{\mu}, \mathbf{Y}) = kV$$
$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\text{JS}}) = \left(k - (k-2)^2 vE\left(\frac{1}{S}\right)\right) V$$

Canonical Examples

MLE for Normal with Both Unknown

By applying JFI, see that the MLE $\hat{\mu} = \bar{X}$ does not depend on σ^2. Plug in so the second term from JFI drops out. Set $t = \sum_{j=1}^{n}(x_j - \overline{x_n})^2$ (it is a constant of the data), take the derivative and find $\hat{\sigma}^2 = \frac{t}{n}$.

House Radiation Levels (Neymann-Scott)

Suppose μ_i be the radiation level at home i and that $Y_{i1}, Y_{i2} \sim \mathcal{N}(\mu_i, \sigma^2)$, all parameters unknown, and all independent.

$$L(\mu_1, \ldots, \mu_n, \sigma^2; \mathbf{y}) = \frac{1}{\sigma^{2n}} e^{\left(-\frac{1}{4\sigma^2}\sum_{j=1}^{n}(y_{j1}-y_{j2})^2 - \frac{1}{\sigma^2}\sum_{j=1}^{n}(\bar{y_j}-\mu_j)\right)}$$

(Solving for MLE) First stage: By inspection, we find that the MLEs are $\hat{\mu}_j = \overline{y_j}$. Second stage: set μ_j's above equal to the MLEs, and *then* take the derivative to find $\hat{\sigma}_2 = \frac{1}{4n}\sum_{j=1}^{n}(y_{j1} - y_{j2})^2$.

(Biasedness) $Y_{j1} - Y_{j2} \sim \mathcal{N}(0, 2\sigma^2)$. $E\hat{\sigma}^2 = \frac{2n\sigma^2}{4n} = \frac{\sigma^2}{2}$. Severe bias! Notice the MLE is off by factor by 2. Could **propose** a new estimator by multiplying by 2.

German Tank Problem

Model: Tank serial #'s $1, 2, \ldots, t$, estimand t. Data: n serial #'s $y_1, \ldots, y_n$. Simple random sample.

$$\begin{aligned} L(t) &= \begin{cases} \frac{1}{\binom{t}{n}}; y_1, \ldots, y_n \in \{1, \ldots, t\} \\ 0 \text{ otherwise} \end{cases} \\ &= \frac{1}{\binom{t}{n}} I(t \geq M) \end{aligned}$$

with $M = \max(y_1, \ldots, y_n)$. By inspection, MLE is found to be $\hat{t} = M$. By the naive def'n of probability, we have:

$$P(M = m) = \frac{\binom{m-1}{n-1}}{\binom{t}{n}}$$

$$EM = \sum_{m=n}^{t} \frac{m\binom{m-1}{n-1}}{\binom{t}{n}} = \frac{n}{n+1}(t+1) \text{ by Feynman Stat 110 HW}$$

We can *propose* a new estimator by fixing the bias, so let $\hat{t} = \frac{n+1}{n}M - 1$ by algebra.
We can find the MOM: $E(Y_j) = \frac{1}{t}\sum_{i=1}^{t} i = \frac{t+1}{2}$. So $\hat{t}_{\text{MOM}} = 2\bar{y} - 1$.