

1

# Random Variables and Distributions

## Universality of Uniform (UoU)

When you plug any CRV into its own CDF, you get a  $\text{Unif}(0,1)$  random variable. When you plug a  $\text{Unif}(0,1)$  r.v. into an inverse CDF, you get an r.v. with that CDF. For example, let's say that a random variable  $X$  has CDF

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

By UoU, if we plug  $X$  into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0,1)$$

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.

## Discrete Distributions

### 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  $Z_1 \perp\!\!\!\perp Z_2$ .

## Continuous Distributions

### 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  $a$ ,  $aX \sim \mathcal{N}(\mu, a^2\sigma^2)$ . 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  $\text{Expo}(\lambda)$ . We know the following:

**CDF**

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

**Expos as a rescaled Expo(1)**

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

**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  $\min(X_1, \dots, X_k) \sim \text{Expo}(\lambda_1 + \lambda_2 + \dots + \lambda_k)$ .

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

### Gamma Distribution

Let us say that  $X$  is distributed  $\text{Gamma}(a, \lambda)$ . We know the following:

**Story** You sit waiting for shooting stars, where the waiting time for a star is distributed  $\text{Expo}(\lambda)$ . You want to see  $n$  shooting stars before you go home. The total waiting time for the  $n$ th shooting star is  $\text{Gamma}(n, \lambda)$ .

**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  $0 \leq k \leq n$ ,

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

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

## $\chi^2$ (Chi-Square) Distribution

Let us say that  $X$  is distributed  $\chi_n^2$ . We know the following:

**Story** A  $\text{Chi-Square}(n)$  is the sum of the squares of  $n$  independent standard Normal r.v.s.

**Properties and Representations**

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

$$X \sim \text{Gamma}(n/2, 1/2)$$

## Multivariate Distributions

### Multinomial Distribution

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

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

**Note** The  $X_1, \dots, X_k$  are dependent.

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

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

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

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

**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,  $\text{Cov}(X_i, X_j) = -np_i p_j$  for  $i \neq j$ .

### Multivariate Normal (MVN) Distribution

A vector  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  is Multivariate Normal if every linear combination is Normally distributed, i.e.,  $t_1 X_1 + t_2 X_2 + \dots + t_k X_k$  is Normal for any constants  $t_1, t_2, \dots, t_k$ . The parameters of the Multivariate Normal are the **mean vector**  $\mu = (\mu_1, \mu_2, \dots, \mu_k)$  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

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

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

## Distribution Properties

### Important CDFs

**Standard Normal**  $\Phi$

**Exponential( $\lambda$ )**  $F(x) = 1 - e^{-\lambda x}$ , 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.  $\chi_n^2 \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$
5.  $\text{NBin}(1, p) \sim \text{Geom}(p)$

## Moments and MGFs

### Moments

### Symmetric Distributions

A distribution is symmetric around its mean  $\mu$  if  $X - \mu$  has the same distribution as  $\mu - X$ . Then for any odd  $n$ , the  $n$ th central moment is  $E(X - \mu)^n$  if it exists.  $X$  is symmetric around  $\mu$  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  $k$ th derivative of the moment generating function, evaluated at 0, is the  $k$ th 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 PDFs and CDFs

### 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**

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

### Independence of Random Variables

Random variables  $X$  and  $Y$  are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of  $Y$  given  $X$  is the marginal distribution of  $Y$

Write  $X \perp\!\!\!\perp Y$  to denote that  $X$  and  $Y$  are independent.

### Multivariate LOTUS

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

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

## Covariance and Transformations

### Covariance and Correlation

**Covariance** is the analog of variance for two random variables.

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

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

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

If  $X_1, X_2, \dots, X_n$  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  $a, b$ :

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

$$\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$$

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

$$\text{Cov}(W + X, Y + Z) = \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) + \text{Cov}(X, Z)$$

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

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

### 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  $\lambda$  arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length  $t$  is  $\text{Pois}(\lambda t)$ .
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  $\lambda$  emails per hour. Let  $T_n$  be the time of arrival of the  $n$ th 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  $T_1 \sim \text{Expo}(\lambda)$ . By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d.  $\text{Expo}(\lambda)$ , i.e., the differences  $T_n - T_{n-1}$  are i.i.d.  $\text{Expo}(\lambda)$ .

## Order Statistics

**Definition** Let's say you have  $n$  i.i.d. r.v.s  $X_1, X_2, \dots, X_n$ . If you arrange them from smallest to largest, the  $i$ th element in that list is the  $i$ th 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  $\leq 42$  and  $X_{(5)}, X_{(6)}, \dots, X_{(n)}$  are  $\geq 42$ .

**Distribution** Taking  $n$  i.i.d. random variables  $X_1, X_2, \dots, X_n$  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  $j$ th order statistic of i.i.d.  $U_1, \dots, U_n \sim \text{Unif}(0, 1)$  is  $U_{(j)} \sim \text{Beta}(j, n - j + 1)$ .

## 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*.

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

**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  $\mu$ . The **sample mean** is

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

. 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  $n \rightarrow \infty$ ,  $\bar{X}_n \rightarrow \mu$  with probability 1. For example, in flips of a coin with probability  $p$  of Heads, let  $X_j$  be the indicator of the  $j$ th 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  $\sim$  to denote *approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable  $Y = X_1 + X_2 + \dots + X_n$  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

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

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

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

#### Asymptotic Distributions using CLT

We use  $\xrightarrow{D}$  to denote *converges in distribution to* as  $n \rightarrow \infty$ . The CLT says that if we standardize the sum  $X_1 + \dots + X_n$  then the distribution of the sum converges to  $\mathcal{N}(0, 1)$  as  $n \rightarrow \infty$ :

$$\frac{1}{\sigma\sqrt{n}}(X_1 + \dots + X_n - n\mu_X) \xrightarrow{D} \mathcal{N}(0, 1)$$

In other words, the CDF of the left-hand side goes to the standard Normal CDF,  $\Phi$ . In terms of the sample mean, the CLT says

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

## 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  $\mathbf{s}Q = \mathbf{s}$ . As a consequence, if  $X_t$  has the stationary distribution, then all future  $X_{t+1}, X_{t+2}, \dots$  also have the stationary distribution. 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 (in which case it's also reversible).

### Random Walk on an Undirected Network

If you have a collection of **nodes**, pairs of which can be connected by undirected **edges**, and a Markov chain is run by going from the current node to a uniformly random node that is connected to it by an edge, then this is a random walk on an undirected network. The stationary distribution of this 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  $(\frac{3}{14}, \frac{3}{14}, \frac{2}{14}, \frac{4}{14}, \frac{2}{14})$ .

## 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  $\mu$  and constant  $a, t > 0$ ,  $P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}$

## Formulas

1. **Geometric Series**  
 $1 + r + r^2 + \dots + r^{n-1} = \sum_{k=0}^{n-1} r^k = \frac{1-r^n}{1-r}$   
 $1 + r + r^2 + \dots = \frac{1}{1-r}$  if  $|r| < 1$
2.  $\mathbf{e}^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$
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**  
 $\int_0^1 x^{t-1} e^{-x} dx = \Gamma(t)$        $\int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$   
Also,  $\Gamma(a+1) = a\Gamma(a)$ , and  $\Gamma(n) = (n-1)!$  if  $n$  is a positive integer.

## 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  $a$ th 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

Let  $f_{\mathbf{Y}}$  be a model for the joint density function of all the observations. Let  $\mathbf{y}$  be the observed value of  $\mathbf{Y}$ . Then the likelihood function is:

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

It is regarded as a function of  $\theta$ , with  $y$  treated as fixed.

**Frequentist Interpretation**  $\theta$  is regarded as fixed but unknown, and it does not have a distribution. If  $L(\theta_2) > L(\theta_1)$ , strictly speaking we cannot say that  $\theta_2$  is more likely than  $\theta_1$  to be the true value of  $\theta$ . However, we can say that  $\theta$  seems more plausible than  $\theta$  as a source for generating the data.

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

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

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  $L(\psi; \mathbf{y}) = L(\theta; \mathbf{y})$ .

## Statistics, Estimands, Estimators, & Estimates

**Statistic** A statistic is a function of  $Y_1, \dots, Y_n$  (and possibly other known quantities). We can write a statistic as  $T(Y)$ , where computing the function  $T$  must not require knowing any unknown parameters. **Note:** The *distribution* of a statistic can depend on unknown parameters.

**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  $\theta$ .

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

## 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  $\theta$  and  $\lambda$ .

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

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

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

## Maximum Likelihood Estimation

The **maximum likelihood estimate** of  $\theta$  is the value  $\hat{\theta}$  that maximizes the likelihood function  $L(\theta; y)$ . The corresponding estimator is called the **maximum likelihood estimator**. That is, if the maximum likelihood estimate is  $T(y)$ , then the maximum likelihood estimator is  $T(Y)$ .

**Regularity conditions** Support must not depend on the value of the estimand. The estimate  $\theta^*$  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  $\theta$ , 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  $\theta$ .

**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  $\theta$  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  $\theta$ . We require that  $L(\hat{\theta}, \hat{\theta}) \geq 0$  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*.

## Estimators for CDF and PDF

### 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:

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

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. The intuitive idea of the KDE for the Gaussian case is to place a little Normal distribution centered at each data point, and then mix all these Normal distributions together to obtain a single smooth PDF. The bandwidth  $h$  controls how tightly concentrated or spread out each of the  $n$  little Normal PDFs is about the  $y$  it is centered at.

## Asymptotics and Information

### Consistency

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

**Sufficient condition** If  $\text{MSE}(\hat{\theta}) \rightarrow 0$  as  $n \rightarrow \infty$ , then  $\hat{\theta}$  is consistent. In particular, if  $\text{Bias}(\hat{\theta}) \rightarrow 0$  and  $\text{Var}(\hat{\theta}) \rightarrow 0$  as  $n \rightarrow \infty$ , 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  $\theta$ , 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,  $\theta^*$  maximizes the expected log-likelihood. the MLE  $\hat{\theta}$  is where the *observed* log-likelihood function has its peak, while  $\theta^*$  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  $s = \frac{\partial l(\theta; y)}{\partial \theta}$ .

As defined, the score function depends on both  $\theta$  and on  $\mathbf{y}$ . In some applications, such as when finding the MLE, we fix  $y$  and look at  $s(\theta; y)$  as a function of  $\theta$ .

In other applications, it turns out to be useful to fix  $\theta$  at its true value  $\theta^*$  and look at the random variables  $(\theta^*; Y)$ . In still other applications, we are interested in some specific hypothesized parameter  $\theta_0$  and we look at the statistics  $(\theta_0; Y)$  (note that  $s(\theta^*; Y)$  is *not* a statistic since  $\theta^*$  is unknown).

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  $\theta$  is:

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

where the subscript of  $\theta$  indicates that we compute the variance under the assumption that the true parameter value is  $\theta$ . 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  $\theta$ . Under regularity conditions,

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

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

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

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

### 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  $\theta$ . Then the MSE of the Rao-Blackwellized estimator  $\hat{\theta}_R B = E(\hat{\theta}|T)$  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

### Definition

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

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

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

## Fun Facts

If  $Y$  is in NEF form (e.g. Normal ( $\sigma^2$  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  $\theta$ .
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:  $\hat{\lambda}|\gamma = \frac{1}{n} \sum y_i^\gamma$