

# Mathematical Statistics I

MAT 5190

Hanan Ather

Fall 2021

## Contents

<b>1</b>	<b>September 8, 2021</b>	<b>4</b>
1.1	Fields and $\sigma$ -Fields . . . . .	4
1.2	Basics Concepts of Probability Theory . . . . .	4
1.3	Conditional Probability and Independence . . . . .	4
1.4	Random Variables . . . . .	5
1.5	Transformation of expectations . . . . .	6
<b>2</b>	<b>September 13, 2021</b>	<b>6</b>
2.1	Product Probability Spaces . . . . .	6
2.2	Discrete Distributions . . . . .	6
<b>3</b>	<b>September 15, 2021</b>	<b>7</b>
3.1	Continuous Distributions . . . . .	7
<b>4</b>	<b>September 20, 2021</b>	<b>8</b>
4.1	Exponential Family . . . . .	8
4.2	Natural Parameterization . . . . .	9
<b>5</b>	<b>September 22, 2021</b>	<b>9</b>
5.1	Location and Scale Families . . . . .	9
5.2	Scale Families . . . . .	10
5.3	Location-scale families . . . . .	11
<b>6</b>	<b>September 27, 2021</b>	<b>11</b>
6.1	Discrete Joint and Marginal Distributions . . . . .	11
6.2	Continuous Joint and Marginal Distributions . . . . .	12
6.3	Multinomial . . . . .	12
<b>7</b>	<b>September 29, 2021</b>	<b>13</b>
7.1	Conditional Distributions (Discrete) . . . . .	13
7.2	Conditional Distributions (Continuous) . . . . .	13
7.3	Independence . . . . .	14
<b>8</b>	<b>October 4, 2021</b>	<b>14</b>
8.1	Bivariate Transformations . . . . .	14

<b>9</b>	<b>October 6, 2021</b>	<b>15</b>
9.1	Hierarchical Models . . . . .	15
<b>10</b>	<b>October 13, 2021</b>	<b>17</b>
10.1	Covariance and Correlation . . . . .	17
10.2	Inequalities . . . . .	17
<b>11</b>	<b>October 18, 2021</b>	<b>18</b>
11.1	Multivariate Distributions . . . . .	18
<b>12</b>	<b>October 20, 2021</b>	<b>19</b>
12.1	Basic Concepts of Random Samples . . . . .	19
<b>13</b>	<b>November 3, 2021</b>	<b>20</b>
13.1	Sampling from Normal Distribution . . . . .	20
13.2	The Derived Distributions: Student's $t$ and Snedecor's $F$ . . . . .	21
<b>14</b>	<b>November 8, 2021</b>	<b>22</b>
14.1	Order Statistics . . . . .	22
<b>15</b>	<b>November 10, 2021</b>	<b>23</b>
15.1	Convergence Concepts . . . . .	23
15.2	Convergence in Probability . . . . .	23
15.3	Almost Sure Convergence and Strong Law of Large Numbers . . . . .	24
15.4	Convergence in Distribution and Central Limit Theorem . . . . .	24
<b>16</b>	<b>November 15, 2021</b>	<b>25</b>
16.1	Inference . . . . .	25
16.2	Sufficiency Principle . . . . .	26
<b>17</b>	<b>November 17, 2021</b>	<b>26</b>
17.1	Factorization Theorem Continued . . . . .	26
<b>18</b>	<b>November 22, 2021</b>	<b>27</b>
18.1	The Sufficiency Principle (Continued) . . . . .	27
18.2	Likelihood Principle . . . . .	28
<b>19</b>	<b>November 24, 2021</b>	<b>29</b>
19.1	Equivariance Principle . . . . .	29
19.2	Methods for finding estimators . . . . .	30
19.3	Methods of Moments . . . . .	30
19.4	Maximum Likelihood Estimators . . . . .	31
<b>20</b>	<b>November 29, 2021</b>	<b>31</b>
20.1	Invariance property of MLE . . . . .	31
20.2	Bayes Estimators . . . . .	32
<b>21</b>	<b>December 1, 2021</b>	<b>33</b>
21.1	Methods of evaluating estimators . . . . .	33
21.2	Best unbiased estimators (UMVUE) . . . . .	33

---

<b>22 December 6, 2021</b>	<b>34</b>
22.1 methods of evaluating estimators (continued) . . . . .	34
22.2 Sufficiency and Unbiasedness . . . . .	34
<b>23 December 8</b>	<b>35</b>

These are course notes for MAT 5190.

## §1 September 8, 2021

### §1.1 Fields and $\sigma$ -Fields

**Definition 1.1** A class (set) of subsets of  $S$  is said to be a field, and is denoted by  $\mathcal{F}$ , if

- (i)  $\mathcal{F}$  is a non-empty class.
- (ii)  $A \in \mathcal{F}$  implies that  $A^c \in \mathcal{F}$  (closed under complementations)
- (iii)  $A_1, A_2, \dots \in \mathcal{F}$  implies that  $A_1 \cup A_2 \in \mathcal{F}$  (that is,  $\mathcal{F}$  is closed under pairwise unions).

Note that there are two key consequences of the definition of a field:

1.  $S, \emptyset \in \mathcal{F}$
2. If  $A_j \in \mathcal{F}, j = 1, 2, \dots, n$ , then,  $\bigcup_{j=1}^n A_j \in \mathcal{F}, \bigcap_{j=1}^n A_j \in \mathcal{F}$  for any finite  $n$ .

**Definition 1.2 (Sigma algebra)** A collection of subsets of  $S$  is called a **sigma algebra** (or **Borel field**), denoted by  $\mathcal{B}$ , if it satisfies the following three properties:

- (a)  $\emptyset \in \mathcal{B}$  (the empty set is an element).
- (b) If  $A \in \mathcal{B}$ , then  $A^c \in \mathcal{B}$  (closed under complement)
- (c) If  $A_1, A_2, \dots \in \mathcal{B}$ , then,  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$  (closed under countable unions).

Many different sigma algebras can be associated with a sample space  $S$ , the collection  $\{\emptyset, S\}$  is the trivial sigma algebra. The only sigma algebra that we will be concerned with is smallest one that contains all open sets in a given sample space.

### §1.2 Basics Concepts of Probability Theory

**Definition 1.3 (Probability Function)** Given a sample space  $S$  and a sigma algebra  $\mathcal{B}$ , a **probability function** is a function  $\mathbb{P}$  with domain  $\mathcal{B}$  that satisfies

1.  $\mathbb{P}(A) \geq 0$  for all  $A \in \mathcal{B}$
2.  $\mathbb{P}(S) = 1$
3. If  $A_1, A_2, \dots \in \mathcal{B}$  are pairwise disjoint, then  $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

### §1.3 Conditional Probability and Independence

**Definition 1.4 (Conditional Probability)** If  $A$  and  $B$  are events in  $S$ , and  $\mathbb{P}(B) > 0$ , then the **conditional** probability of  $A$  given  $B$  is

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

Note that in the calculation of conditional probability the event  $B$  becomes the sample space. The intuition is that our original sample space  $S$  has been updated to  $B$ . All further occurrences are then calibrated with respect to their relation with  $B$ .

**Theorem 1.5 (Bayes' Rule)** —  $A_1, A_2, \dots$  be a partition of the sample space, and let  $B$  be any set. Then for each  $i = 1, 2, \dots$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

**Definition 1.6 (Independent, Mutually independent)** Two events  $A$  and  $B$  are **statistically independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

A collection of events  $A_1, \dots, A_n$  are **mutually independent** if for any subcollection  $A_{i_1}, \dots, A_{i_k}$ , we have

$$\mathbb{P}\left(\bigcup_{j=1}^k A_{ij}\right) = \prod_{j=1}^k \mathbb{P}(A_{ij})$$

## §1.4 Random Variables

Given a probability space  $(S, \mathcal{F}, \mathbb{P})$ , the main objective of probability theory is that of calculating probabilities of events which may be of importance to us. The sample space  $S$  may be quite an abstract set, thus we can facilitate our calculations by a transformation of the sample space  $S$ , into a subset of the real line  $\mathbb{R}$ . This is achieved by a **random variable**<sup>1</sup> which is a function from sample space  $S$  into  $\mathbb{R}$ . With every random variable  $X$  we associate a function called the cumulative distribution function of  $X$ .

**Definition 1.7 (cdf, pmf, pdf)** The **cumulative distribution function** (cdf) of a random variable  $X$ , denoted by  $F_X(x)$ , is defined by

$$F_X(x) = \mathbb{P}_X(X \leq x), \text{ for all } x.$$

The **probability mass function** (pmf) of a *discrete* random variable  $X$  is given by

$$f_X(x) = \mathbb{P}_X(X = x) \text{ for all } x.$$

The **probability density function** (pdf),  $f_X(x)$  of a *continuous* random variable is  $X$  is the function that satisfies

$$F_X(x) = \int_{-\infty}^x f_X(t)dt \text{ for all } x$$

**Fact 1.1** (Valid CDFs). Any CDF  $F$  has the following properties:

- Right-continuous: that is, for any  $a$ , we have

$$F(a) = \lim_{x \rightarrow a^+} F(x)$$

<sup>1</sup>A random variable is a real-valued function with a domain  $\Omega$  which has an extra property called **measurability**.

- Convergence to 0 or 1 in the limits:

$$\lim_{x \rightarrow -\infty} F(x) = 0 \quad \lim_{x \rightarrow \infty} F(x) = 1$$

## §1.5 Transformation of expectations

**Theorem 1.8** — Let  $X$  have a cdf  $F_X(x)$ , let  $Y = g(X)$ , and let  $\mathcal{X}$  and  $\mathcal{Y}$  be respective sample spaces.

- (a) if  $g$  is an increasing function on  $\mathcal{X}$ ,  $F_Y(y) = F_X(g^{-1}(y))$  for  $y \in \mathcal{Y}$ .
- (b) if  $g$  is a decreasing function on  $\mathcal{X}$  and  $X$  is continuous random variable,  $F_Y(y) = 1 - F_X(g^{-1}(y))$

**Theorem 1.9** — Let  $X$  have a pdf  $f_X(x)$  and let  $Y = g(X)$ ,  $g$  is monotone function. Let  $\mathcal{X}$  and  $\mathcal{Y}$  be the respective sample spaces. Suppose  $f_X(x)$  is continuous on  $\mathcal{X}$  and that  $g^{-1}(y)$  has continuous derivative on  $\mathcal{Y}$ . Then the pdf of  $Y$  is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{if } y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

**Theorem 1.10 (Probability integral transformation)** — Let  $X$  have a continuous cdf  $F_X(x)$  and define the random variable  $Y$  as  $Y = F_X(X)$ . Then  $Y$  is uniformly distributed on  $(0, 1)$ , that is,  $\mathbb{P}(Y \leq y) = y, 0 < y < 1$ .

## §2 September 13, 2021

### §2.1 Product Probability Spaces

If we consider experiments  $\mathcal{E}_1$  and  $\mathcal{E}_2$  with respective probability spaces  $(S_1, \mathcal{F}_1, \mathbb{P}_1)$  and  $(S_2, \mathcal{F}_2, \mathbb{P}_2)$ , then  $(\mathcal{E}_1, \mathcal{E}_2) = \mathcal{E}_1 \times \mathcal{E}_2$  has the sample space  $S = S_1 \times S_2$ . The appropriate  $\sigma$ -field  $\mathcal{F}$  of events in  $S$  is defined by first defining the class  $\mathcal{C}$ :

$$\mathcal{C} = \{\mathcal{F}_1 \times \mathcal{F}_2 : F_1 \in \mathcal{F}_1, F_2 \in \mathcal{F}_2\}$$

$$\text{where, } \mathcal{F}_1 \times \mathcal{F}_2 = \{(s_1, s_2) : s_1 \in A_1, s_2 \in A_2\}$$

### §2.2 Discrete Distributions

**Discrete uniform distribution.** We consider an urn with  $N$  balls, numbered  $1, \dots, N$ . Let  $X$  be the number of the randomly select ball from the urn. We have

$$\mathbb{P}(X = x|N) = \frac{1}{N}, x = 1, \dots, N.$$

We say that  $X$  is *discrete uniform distribution* and we write  $X \sim \text{Discrete Uniform}(1, N)$ . We can prove that  $\mathbb{E}(X) = \frac{N+1}{2}$  and  $\mathbb{V}(X) = \frac{(N+1)(N-1)}{12}$ . More generally, we have  $X \sim \text{Discrete Uniform}(N_0, N_1)$  if

$$\mathbb{P}(X = x|N) = \frac{1}{N_1 - N_0 + 1}, x = N_0, \dots, N_1.$$

**Hypergeometric distribution.** Consider an urn which contains  $N$  balls:  $M$  red and  $N - M$  green. Select a sample size of  $K$ . Let  $X$  be a random variable which gives the total number of red balls in this sample. We say that  $X$  has a *hypergeometric distribution* and we write  $X \sim \text{Hypergeometric}(N, M, K)$ .

$$\mathbb{P}(X = x | N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, x = 0, \dots, K; M - (N - K) \leq x \leq M$$

**Poisson distribution.** Let  $\lambda$  be the average number of events which occur in a fixed interval of time and  $X$  be the random number of events which occur in the same interval. We have

$$\mathbb{P}(X = x | \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, x = 0, 1, 2, \dots$$

**Negative Binomial distribution.** Consider a sequence of identical Bernoulli trials with probability  $p$  of success. Let  $X$  be the number of trials required to get fixed number  $r$  successes. A combinatorial argument shows that

$$\mathbb{P}(X = x | r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, x = r, r+1, \dots$$

## §3 September 15, 2021

### §3.1 Continuous Distributions

A random variable  $X$  has a *continuous distribution* if its cdf is continuous. Today's lecture covered some of the most commonly used continuous distributions.

**Uniform Distribution.** This is a continuous distribution with pdf:

$$f(x|a, b) = \frac{1}{b-a}, \quad a \leq x \leq b, \quad \text{where } -\infty < a < b < \infty$$

**Gamma Distribution.** This is a continuous distribution with pdf:

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty.$$

where  $0 < \alpha, \beta < \infty$  and  $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$  is the *gamma function*. We write  $X \sim \text{Gamma}(\alpha, \beta)$ . Note that

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha), \quad \Gamma(1) = 1, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

**Beta Distribution.** This is a continuous distribution with pdf:

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1, \text{ where } 0 < \alpha, \beta < \infty.$$

Note that  $B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$  is the *beta function*. We write  $X \sim \text{Beta}(\alpha, \beta)$ . We have

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

**Cauchy Distribution.** This is a continuous distribution with the pdf:

$$f(x|\theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}, \quad -\infty < x < \infty$$

Where  $-\infty < \theta < \infty$ . This is a “pathological” example, since  $\mathbb{E}X = \infty$ .

**Lognormal Distribution.** This is a continuous distribution with the following pdf:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \frac{1}{x} \exp \left\{ -\frac{(\log x - \mu)^2}{2\sigma} \right\}, \quad 0 < x < \infty$$

where  $-\infty < \mu < \infty$ ,  $0 < \sigma < \infty$ . We write  $X \sim \text{Lognormal}(\mu, \sigma^2)$ .

**Double Exponential Distribution.** This is a continuous distribution with pdf:

$$f(x|\mu, \sigma) = \frac{1}{2\sigma} \exp \left\{ -\frac{|x - \mu|}{\sigma} \right\}, \quad -\infty < x < \infty$$

Where  $-\infty < \mu < \infty$ ,  $0 < \sigma < \infty$ . We write  $X \sim \text{Double Exponential}(\mu, \sigma)$ . Its pretty straight forward to check that

$$\mathbb{E}X = \mu, \quad \mathbb{V}(X) = 2\sigma^2.$$

## §4 September 20, 2021

### §4.1 Exponential Family

A family of pdf's (or pmf's) is called *exponential family* if it can be written in the form

$$f(x|\theta) = h(x)c(\theta) \exp \left\{ \sum_{i=1}^k t_i(x)w_i(\theta) \right\} \quad (1)$$

where  $h(x) \geq 0, c(\theta) \geq 0, w_i(\theta) \in \mathbb{R}, t_i(x) \in \mathbb{R}$ . Here the parameter  $\theta = \theta_1, \dots, \theta_d$  is vector-valued. If  $d = k$ , the family is called a *full exponential family*. If  $d < k$ , the family is called a *curved exponential family*.

#### Binomial family with $n$ known:

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \left( \frac{p}{1-p} \right)^x = \binom{n}{x} (1-p)^n \exp \left\{ x \log \left( \frac{p}{1-p} \right) \right\} \\ &= h(x)c(\theta) \exp \left\{ \sum_{i=1}^k t_i(x)w_i(\theta) \right\} \end{aligned}$$

#### Poisson Family:

$$f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{1}{x!} e^{-\lambda} \exp\{x \log \lambda\} = h(x)c(\theta) \exp \left\{ \sum_{i=1}^k t_i(x)w_i(\theta) \right\}$$



**Theorem 4.1** — If  $X$  is a random variable whose probability density function is given by (1) then

$$\mathbb{E} \left( \sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right) = -\frac{\partial}{\partial_j} \log c(\theta)$$

$$\mathbb{V}ar \left( \sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\theta) - \mathbb{E} \left( \sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_j^2} t_i(X) \right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\theta)$$

## §4.2 Natural Parameterization

If we use  $\eta_i = w_i(\theta)$  in formula (1) and  $\eta = (\eta_1, \dots, \eta_k)$ , we obtain the *natural parametrization*:

$$f(x|\eta) = h(x)c^*(\eta)\exp \left\{ \sum_{i=1}^k t_i(x)\eta_i \right\} \quad (2)$$

where  $h(x)$  and  $t_i(x)$  are the same as in formula (2). The natural space is  $\mathcal{H} = \{\eta : \int_{-\infty}^{\infty} \exp\{\sum_{i=1}^k t_i(x)\eta_i\} dx < \infty\}$ . We have

$$c^*(\eta) = \frac{1}{\int_{-\infty}^{\infty} \exp\{\sum_{i=1}^k t_i(x)\eta_i\} dx}, \quad \eta \in \mathcal{H}$$

**Normal Family.** The natural parametrization if  $\eta_1 = 1/\sigma^2$ ,  $\eta_2 = \mu/\sigma^2$  with natural parameter space  $\mathcal{H} = \{(\eta_1, \eta_2) : 0 < \eta_1 < \infty, -\infty < \eta_2 < \infty\}$ . We have  $\mu = \eta_2/\eta_1$  and  $\sigma^2 = 1/\eta_1$  and hence the normal pdf

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} \exp \left\{ -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} \right\}$$

can be written as,

$$f(x|\eta_1, \eta_2) = \frac{\sqrt{\eta_1}}{\sqrt{2\pi}} \exp \left\{ -\frac{\eta_2^2}{2\eta_1} \right\} \exp \left\{ -\frac{\eta_1 x^2}{2} + \eta_2 x \right\}$$

$$= f(x|\eta) = h(x)c^*(\eta)\exp \left\{ \sum_{i=1}^k t_i(x)\eta_i \right\}$$

## §5 September 22, 2021

### §5.1 Location and Scale Families

In this lecture we discussed three techniques for constructing families of distributions. Each of these technique relies on first specifying a single pdf,  $f(z)$ , called the *standard pdf* for the family. The other pdf's in the family are generated by applying a certain transformation to the standard pdf.

**Theorem 5.1** — Let  $f(z)$  be any odf and  $\mu$  and  $\sigma > 0$  be arbitrary constants. Then the following function is also a pdf:

$$g(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)$$

The family  $\{f(x - \mu) : -\infty < \mu < \infty\}$  is called a *location family* with a standard pdf  $f(z)$ . The location parameter  $\mu$  shifts the graph  $f(z)$  with  $\mu$ , without changing its shape.

**Representation:** Let  $Z$  be a random variable with pdf  $f(z)$  and  $X = \mu + Z$ . Then the pdf of  $X$  is  $g(x|\mu)$ . If  $F_z(z)$  and  $F_x(x)$  are cdf's of  $Z$ , respectively  $X$ , then

$$F_X(x) = F_z(x - \mu).$$

**Uniform location family.** By taking  $f(z)$  be a uniform  $U(a, b)$  pdf, we generate the following location family:

$$g(x|\mu) = \frac{1}{b - a}, \quad a + \mu < x < b + \mu$$

**Exponential location family.** By taking  $f(z)$  to be Exponential( $\beta$ ) pdf ( $\beta$  is a fixed value), we generate the following location family

$$g(x|\mu) = \frac{1}{\beta} e^{-(x-\mu)/\beta}, \quad \mu < x < \infty$$

## §5.2 Scale Families

The family  $\{(1/\sigma)f(x/\sigma) : 0 < \sigma < \infty\}$  is called a *scale family* with standard pdf  $f(z)$ .

A scale parameter  $\sigma > 1$  stretches the graph with pdf  $f(z)$  without changing its basic shape. Similarly a scale parameter  $\sigma < 1$  contracts the graph of  $f(z)$ .

**Representation:** Let  $Z$  be a random variable with  $f(z)$  and  $X = \sigma Z$ . Then the pdf of  $X$  is  $g(x|\sigma) = (1/\sigma)f(x/\sigma)$ . If  $F_Z(z)$  and  $F_x(x)$  are the cdf's of  $Z$  and  $X$ , respectively, then

$$F_X(x) = F_Z\left(\frac{x}{\sigma}\right).$$

**Gamma Scale Family.** By taking  $f(z)$  be the pdf of Gamma( $\alpha, 1$ ), (where  $\alpha$  is fixed), we generate the following scale family

$$g(x|\sigma) = \frac{1}{\Gamma(\alpha)\sigma^\alpha} x^{\alpha-1} e^{-x/\sigma}, \quad 0 < x < \infty$$

**Double Exponential.** By taking  $f(z)$  as the pdf of Double Exponential( $0, \sigma$ ), we generate the following scale family

$$g(x|\sigma) = \frac{1}{2\sigma} e^{-|x|/\sigma}, \quad -\infty < x < \infty$$

### §5.3 Location-scale families

The following family of distributions  $\{(1/\sigma)f(x - \mu)/\sigma) : -\infty < \mu < \infty, 0 < \sigma < \infty\}$  is called *location-scale family* with standard pdf  $f(z)$ .

Let  $Z$  be a random variable with pdf  $f(z)$  and  $X = \mu + \sigma Z$ . Then pdf of random variable  $X$  is given by  $(1/\sigma)f(x - \mu)/\sigma$ . If  $F_Z(z)$  and  $F_X(x)$  are the cdf's of  $Z$  and  $X$ , respectively, then

$$F_X(x) = F_Z\left(\frac{x - \mu}{\sigma}\right)$$

Therefore, clearly we have

$$\mathbb{E}X = \mu + \sigma\mathbb{E}Z, \quad \text{Var}X = \sigma^2\text{Var}Z$$

Generally, the standard pdf  $f(z)$  is chosen in such a way that  $\mathbb{E}Z = 0$  and  $\text{Var}Z = 1$ , this results in  $\mathbb{E}X = \mu$  and  $\text{Var}X = \sigma$ .

## §6 September 27, 2021

In this section we are going to consider events that *co-occur*, and revisit concepts such as *independence* and *conditional probability*. We will learn how to handle random variables that co-occur.

### §6.1 Discrete Joint and Marginal Distributions

**Definition 6.1 (Joint PMF)** Let  $(X, Y)$  be a bi-variate random vector. We say that the distribution of  $(X, Y)$  is *discrete* if the possible values of  $(X, Y)$  are countable. In this case the function,

$$f(x, y) = f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

is called the **joint pmf** of  $X, Y$ .

What is the most important information about a random variable? The PMF, or the PDF. The multivariate analogue to the **Joint function**, which takes in a value of two or more random variables, and returns the probability that those two variables jointly take on those values.

$$\mathbb{P}(X = x, Y = y) \quad \text{Joint Probability of } X \text{ and } Y$$

Should be read as: “Probability  $X$  takes on the value  $x$  and  $Y$  takes on the value  $y$ ”.

A joint probability table is a way of specifying the “joint” probability distribution between multiple random variables. It does so by keeping a multi-dimensional lookup table, so essentially any assignment of the random variables,  $\mathbb{P}(X = x, Y = y, \dots)$  can be directly looked up. A probability mass table is a brute force way to store the joint probabilities of random variables.

**Property 1.** If  $A$  is a subset of  $\mathbb{R}^2$

$$\mathbb{P}((X, Y) \in A) = \sum_{(x, y) \in A} \mathbb{P}(X = x, Y = y)$$

**Property 2.** If  $g(x, y)$  is a real-valued function, then

$$\mathbb{E}g(X, Y) = \sum_{x, y} g(x, y) f(x, y)$$

**Definition 6.2 (Marginal pmf)** Let  $f(x, y)$  be the joint PMF of the discrete random vector  $(X, Y)$ , the PMF of  $X$  is called the **marginal pmf** of  $X$ , denoted by  $f_X(x)$ . Similarly, the PMF of  $Y$  is called the marginal PMF of  $Y$ , denoted by  $f_Y(y)$ .

The marginals can be computed by the following formulas:

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y), \quad f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y)$$

## §6.2 Continuous Joint and Marginal Distributions

**Definition 6.3** Let  $(X, Y)$  be a bivariate random vector. We say the distribution of  $(X, Y)$  is *continuous* if the joint CDF of  $(X, Y)$  is defined by  $F(u, v) = \mathbb{P}(X \leq u, Y \leq v)$  is continuous. In this case the function  $f(x, y)$  which satisfies the condition

$$F(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$$

is called the **joint pdf** of  $(X, Y)$ .

Note that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1, \quad f(x, y) \geq 0 \text{ for all } x, y$$

**Property 1.** If  $A$  is a subset of  $\mathbb{R}^2$

$$\mathbb{P}((X, Y) \in A) = \int_A \int f(x, y) dx dy$$

**Property 2.** If  $g(x, y)$  is a real-valued function, then

$$\mathbb{E}g(X, Y) = \int_A \int g(x, y) f(x, y) dx dy$$

The marginals can be computed by the following formulas:

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

## §6.3 Multinomial

The multi-nomial distribution is a **parametric** distribution for multiple random variables.

## §7 September 29, 2021

### §7.1 Conditional Distributions (Discrete)

Let  $(X, Y)$  be discrete bivariate random vector with joint pmf  $f(x, y)$  and marginal pmf's  $f_X(x)$  and  $f_Y(y)$ . For all fixed  $x \in \mathcal{X} = \{x : f_X(x) > 0\}$ , we define

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

The function  $y \mapsto f(y|x)$  is called **conditional pmf** of  $Y$  given  $X = x$ . For every  $x \in \mathcal{X}$ , the function  $y \mapsto f(y|x)$  is a pmf, since

$$f(y|x) \geq 0 \text{ for all } y, \quad \text{and} \quad \sum_y f(y|x) = 1.$$

**Property 1.** For every fixed  $x \in \mathcal{X}$  and for every set  $A$

$$\mathbb{P}(Y \in A | X = x) = \sum_{y \in A} f(y|x).$$

The function  $y \mapsto F(y|x) := \mathbb{P}(Y \leq y | X = x)$  is called the **conditional cdf** of  $Y$  given that  $X = x$ .

**Property 2.** For every fixed  $x \in \mathcal{X}$  and for every real-valued function  $g$

$$\mathbb{E}(g(Y) | X = x) = \sum_y g(y) f(y|x).$$

### §7.2 Conditional Distributions (Continuous)

Let  $(X, Y)$  be continuous random vector with joint pdf  $f(x, y)$  and marginal pdf's  $f_X(x)$  and  $f_Y(y)$ . For all fixed  $x \in \mathcal{X} = \{x : f_X(x) > 0\}$ , we define

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

The function  $y \mapsto f(y|x)$  is called **conditional pdf** of  $Y$  given  $X = x$ . For every  $x \in \mathcal{X}$ , the function  $y \mapsto f(y|x)$  is a pdf, since

$$f(y|x) \geq 0 \text{ for all } y, \quad \text{and} \quad \int_{-\infty}^{\infty} f(y|x) dy = 1.$$

**Note that in the continuous case  $f(y|x) \neq \mathbb{P}(Y = y | X = x)$ !**

**Notation 1.** For any fixed  $x \in \mathcal{X}$  and for every set  $A$

$$\mathbb{P}(Y \in A | X = x) = \int_A f(y|x) dy$$

The function  $y \mapsto F(y|x) := \mathbb{P}(Y \leq y | X = x)$  is called the **conditional cdf** of  $Y$  given that  $X = x$ .

**Notation 2.** For every fixed  $x \in \mathcal{X}$  and for every real-valued function  $g$

$$\mathbb{E}(g(Y) | X = x) = \int_{-\infty}^{\infty} g(y) f(y|x) dy.$$

## §7.3 Independence

From the definition of conditional pmf we can derive the joint pdf (or pmf) of  $(X, Y)$  as

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x).$$

**Definition 7.1 (Independent R.V's)** Let  $(X, Y)$  be a random vector with joint pdf or pmf  $f_{X,Y}(x, y)$ , and marginal pdf's or pmf's  $f_X(x)$  and  $f_Y(y)$ . If

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \quad \text{for all } x, y \in \mathbb{R}$$

and we say that  $X$  and  $Y$  are *independent*.

**Lemma 7.2** — Let  $(X, Y)$  be a bi-variate vector with joint pdf or pmf  $f(x, y)$ . The variables  $X$  and  $Y$  are independent if and only if there exists some function  $g(x), h(y)$  such that

$$f(x, y) = g(x)h(y), \quad x \in \mathbb{R}, y \in \mathbb{R}.$$

Note that function  $g(x), h(y)$  “coincide” with marginal pdf's  $f_X(x)$  and  $f_Y(y)$  up to a constant, i.e, there exists some positive constants  $C_1, C_2$  with  $C_1 \times C_2 = 1$  such that

$$g(x) = C_1 f_X(x), \quad h(y) = C_2 f_Y(y).$$

**Theorem 7.3** — If  $X$  and  $Y$  are independent, then for any function  $\varphi(x), \psi(y)$

$$\mathbb{E}[\varphi(X)\psi(Y)] = \mathbb{E}[\varphi(X)]\mathbb{E}[\psi(Y)]$$

and,

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

## §8 October 4, 2021

### §8.1 Bivariate Transformations

In this lecture our focus will be on computing the pdf of a bivariate random vector  $(U, V)$  defined by

$$U = g_1(X, Y), \quad V = g_2(X, Y)$$

where  $(X, Y)$  is a random vector with a known joint pdf  $f(x, y)$  and  $g_1, g_2$  are defined functions.

**Theorem 8.1** — Let  $(X, Y)$  be a bivariate random vector with joint pdf  $f_{X,Y}(x, y)$  and  $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ . Let  $g = (g_1, g_2) : \mathcal{A} \rightarrow \mathcal{B}$  be a one-to-one transformation. Denote  $g^{-1} := h = (h_1, h_2)$  and let  $J$  be the Jacobian of the transformation

$$J = \begin{bmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{bmatrix}$$

Then the joint pdf of  $(U, V)$  is

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \cdot |J|, \quad (u, v) \in \mathcal{B}.$$

If the transformation  $g$  is not one-to-one on  $\mathcal{A}$ , but we can find a partition  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$  such that  $g = g^{(i)} : \mathcal{A}_i \rightarrow \mathcal{B}_i$  is one-to-one for each  $i = 1, \dots, k$ , then we can still write down a formula for the joint pdf of  $(U, V) = g(X, Y)$ :

$$f_{U,V}(u, v) = \sum_{i=1}^j f_{X,Y}(h_1^{(i)}(u, v), h_2^{(i)}(u, v)) \cdot |J_i| \mathbb{I}_{(u,v) \in \mathcal{B}_i}$$

where  $h^{(i)} = (h_1^{(i)}, h_2^{(i)})$  are the inverse of  $g^{(i)}$  and  $J_i$  is the corresponding Jacobian.

## §9 October 6, 2021

### §9.1 Hierarchical Models

Thus far we have seen random variables which have single distributions, possibly depending on parameters. However, we can think of the parameter of a distribution as being a random variable, which itself has a distribution.

**Example 9.1** (Binomial-Poisson hierarchy) — Classic example of hierarchical model is the following: An insect lays a large number of eggs, each surviving with probability  $p$ . On average, how many eggs will survive? The “large number” of eggs laid is a random variable, often taken to be  $\text{Poisson}(\lambda)$ . Furthermore, if we assume that each eggs survival is independent, we have Bernoulli trials. Therefore if  $X$  = number of survivors and  $Y$  = number of eggs laid, we have

$$X|Y \sim \text{binomial}(Y, p)$$

$$Y \sim \text{Poisson}(\lambda),$$

a hierarchical model.

**Solution:**

The random variable of interest,  $X = \text{number of survivors}$ , has the distribution given by

$$\begin{aligned}
 \mathbb{P}(X = x) &= \sum_{y=0}^{\infty} \mathbb{P}(X = x, Y = y) \\
 &= \sum_{y=0}^{\infty} \mathbb{P}(X = x|Y = y) \mathbb{P}(Y = y) \\
 &= \sum_{y=x}^{\infty} \left[ \binom{y}{x} p^x (1-p)^{y-x} \right] \left[ \frac{e^{-\lambda \lambda^y}}{y!} \right] \\
 &\quad \vdots \text{ (after some algebraic simplifications)} \\
 &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda}
 \end{aligned}$$

Thus,  $X \sim \text{Poisson}(\lambda p)$ . Thus, any marginal inference on  $X$  is with respect to a  $\text{Poisson}(\lambda p)$  distribution, with  $Y$  playing no part at all. Introduction  $Y$  in the hierarchy was mainly to aid our understanding the model.

Now we can easily compute the expected value

$$\mathbb{E}[X] = \lambda p$$

so, on average,  $\lambda p$  eggs will survive.

Sometimes calculations can be greatly simplified using the following theorem. Recall that  $\mathbb{E}(X|y)$  is a function of  $y$  and  $\mathbb{E}(X|Y)$  is a random variable whose distribution depends on the value of  $Y$ .

**Theorem 9.2** — If  $X$  and  $Y$  are random variable, then

$$\mathbb{E}(X) = \mathbb{E}\mathbb{E}(X|Y)$$

and,

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

*Proof.* By definition

$$\underbrace{\mathbb{E}(X|Y) = \sum_x x \mathbb{P}(X = x|Y = y)}_{\text{Average of } X \text{ when we fix } Y = y}$$

But  $Y$  is a random variable, so if average over all realizations of  $Y$ , we have,

$$\mathbb{E}_Y(\mathbb{E}_X(X|Y)) = \sum_y \underbrace{\sum_x x \mathbb{P}(X = x|Y = y)}_{\mathbb{E}(X|Y)} \cdot \mathbb{P}(Y = y)$$

by the definition of joint density, we can re-write the equation,

$$\implies \sum_y \sum_x x \mathbb{P}(x, y) = \sum_x \sum_y x \mathbb{P}(x, y) = \sum_x x \underbrace{\sum_y \mathbb{P}(x, y)}_{\mathbb{P}(X=x)} = \sum_x x \mathbb{P}(X = x) = \mathbb{E}(X).$$

□



## §10 October 13, 2021

### §10.1 Covariance and Correlation

**Definition 10.1 (Covariance and Correlation)** Let  $X$  and  $Y$  be random variables such that  $\mathbb{E}X^2 < \infty$ ,  $\mathbb{E}Y^2 < \infty$ . The **covariance** of  $X$  and  $Y$  is defined by

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$$

The **correlation** is defined as

$$\rho_{X,Y} := \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}\text{ar}(X)\mathbb{V}\text{ar}(Y)}}$$

If  $X$  and  $Y$  are independent random variables, then  $\text{Cov}(X, Y) = 0$ . The converse is not true **in general**. However, the converse is true if  $(X, Y)$  are **bivariate normal** distribution.

Furthermore, if  $X$  and  $Y$  are random variables and  $a$  and  $b$  are constants, then

$$\mathbb{V}\text{ar}(aX, bY) = a^2\mathbb{V}\text{ar}(X) + b^2\mathbb{V}\text{ar}(Y) + 2ab\text{Cov}(X, Y).$$

**Consequence:** If  $X$  and  $Y$  are positively correlated (i.e,  $\text{Cov}(X, Y) \geq 0$ ), then

$$\mathbb{V}\text{ar}(X + Y) \geq \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)$$

**Note:** A special case of positive dependence structure “**association**”. We say that  $X$  and  $Y$  are *weakly associated* if

$$\text{Cov}(g(X), h(Y)) \geq 0$$

for any non-decreasing functions  $g, h$  for which covariance exists.

### §10.2 Inequalities

The inequalities below, although often stated in terms of expectation, rely mainly on properties of numbers. They are all based on the following simple lemma.

**Lemma 10.2** — Let  $a$  and  $b$  be any positive numbers, and let  $p$  and  $q$  any positive numbers (greater than 1) stratifying

$$\frac{1}{p} + \frac{1}{q} = 1$$

Then,

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality of  $a^p = b^q$ .

*Proof.* Fix,  $b$  and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

To minimize  $g(a)$ , we differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \implies a^{p-1} - b = 0 \implies a = b^{1/(p-1)}$$

We can also check the second derivative to establish that this is indeed a minimum. The value of the function at the minimum is 0.  $\square$

**Theorem 10.3 (Holder's Inequality)** — Let  $X$  and  $Y$  be any two random variables, and let  $p$  and  $q$  satisfy. Then

$$|\mathbb{E}XY| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$$

*Proof.* The first inequality follows from the fact that  $-|XY| \leq XY \leq |XY|$  and theorem 2.2.5 in the textbook. To prove second inequality, define,

$$a = \frac{|X|}{(\mathbb{E}|X|^p)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(\mathbb{E}|Y|^q)^{1/q}}$$

Applying the lemma above, and take the expectation of both side. The expectation of left-hand side is 1 and rearranging gives us the second inequality.  $\square$

Perhaps the most famous special case of Holder's Inequality is the Cauchy-Schwartz ( $p = 2$ ).

**Theorem 10.4 (Cauchy-Schwartz Inequality)** — For any two random variables  $X$  and  $Y$ ,

$$|\mathbb{E}XY| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^2)^{1/2}(\mathbb{E}|Y|^2)^{1/2}$$

## §11 October 18, 2021

### §11.1 Multivariate Distributions

The random vector  $\mathbf{X} = (X_1, \dots, X_n)$  has a sample space in  $\mathbb{R}^n$ . If  $(X_1, \dots, X_n)$  is discrete random vector (the sample space is countable), then the *joint pmf* of  $(X_1, \dots, X_n)$  is the function define by  $f(\mathbf{x}) = f(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$  for each  $(x_1, \dots, x_n)$ . Then for any  $A \in \mathbb{R}^n$ ,

$$\mathbb{P}(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$$

If  $(X_1, \dots, X_n)$  is continuous random vector, the joint pdf of  $(X_1, \dots, X_n)$  is a function  $f(x_1, \dots, x_n)$  that satisfies

$$\mathbb{P}(\mathbf{x}) = \int \cdots \int_A f(\mathbf{x}) d\mathbf{x} = \int \cdots \int_A f(x_1 \dots x_n) dx_1 \dots dx_n$$

These are  $n$ -fold integrals with limits of integration set so that the integration is over all points of  $\mathbf{x} \in A$ .

**Marginal pdf or pmf** of any subset of of the coordinates of  $(X_1, \dots, X_n)$  can

be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates.

$$f(x_1, \dots, x_k) = \int_{\mathbb{R}^{n-k}} f(x_1, \dots, x_n) dx_{k+1} \dots dx_n = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_{k+1} \dots dx_n$$

or in the discrete case

$$f(x_1, \dots, x_k) = \sum_{x_{k+1}, \dots, x_n \in \mathbb{R}^{n-k}} f(x_1, \dots, x_n),$$

for every  $(x_1, \dots, x_n) \in \mathbb{R}^n$ .

## §12 October 20, 2021

### §12.1 Basic Concepts of Random Samples

**Definition 12.1** (Random sample) The random variables  $X_1, \dots, X_n$  are called a **random sample of size n** from population  $f(x)$  if  $X_1, \dots, X_n$  are mutually independent random variables and the marginal pdf or pmf of each  $X_i$  is the same function  $f(x)$

The joint pdf of the random sample is given by

$$f(x_1, \dots, x_n) = f(x_1) \dots f(x_n) = \prod_{i=1}^n f(x_i)$$

The joint pdf can be used to calculate the probabilities involving samples. If the population pdf is a member of the parametric family with a pdf or pmf given by  $f(x|\theta)$ , then the joint pdf or pmf is

$$f(x_1, \dots, x_n|\theta) = f(x_1|\theta) \dots f(x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

**Definition 12.2** Let  $X_1, \dots, X_n$  be a random sample and  $T(x_1, \dots, x_n)$  be real vector valued function, whose domain includes the sample space of  $X_1, \dots, X_n$ . The random variable (or vector)  $Y = T(X_1, \dots, X_n)$  is called a **statistic**.

The definition of a statistics is very broad, with only one restriction being that a statistic cannot be a function of the parameter. The following three statistics are used to provide sample summaries for data :

1.  $\bar{X} = \frac{X_1 + \dots + X_n}{n}$ , the sample mean.
2.  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , the sample variance
3.  $S = \sqrt{S^2}$ , standard deviation.

The next result is useful for studying sampling distributions.

**Lemma 12.3** — Let  $X_1, \dots, X_n$  be a random sample from a population and let  $g(x)$  be a function such that  $\mathbb{E}g(X_1)$  and  $\text{Var}[g(X_1)]$  exist. Then

$$\mathbb{E} \left( \sum_{i=1}^n g(X_i) \right) = n(\mathbb{E}g(X_1))$$

and

$$\text{Var} \left( \sum_{i=1}^n g(X_i) \right) = n(\text{Var}g(X_1))$$

*Proof.* Details of the proof are in lecture notes and textbook. □

Let  $X_1, \dots, X_n$  be a random sample from a population with mean  $\mu$  and variance  $\sigma^2 < \infty$ . Then

1.  $\mathbb{E}\bar{X} = \mu$
2.  $\text{Var}\bar{X} = \frac{\sigma^2}{n}$
3.  $\mathbb{E}S^2 = \sigma^2$

**Theorem 12.4** — Let  $X_1, \dots, X_n$  be a random sample from a population with a mgf  $M_X(t)$ . Then the mgf of  $\bar{X}$  is given by

$$M_{\bar{X}}(t) = \left[ M_X\left(\frac{t}{n}\right) \right]^n$$

If the mgf does not exist or doesn't have a closed form, we can use the following result to derive the distribution of  $\bar{X}$ .

**Theorem 12.5** — Let  $X$  and  $Y$  be independent continuous random variables with pdf's  $f_X(x)$  and  $f_Y(y)$ . Then the pdf of  $U = X + Y$  is given by

$$f_U(u) = \int_{-\infty}^{\infty} f_X(x)f_Y(u-x)dx = \int_{-\infty}^{\infty} f_X(u-y)f_Y(y)dy$$

## §13 November 3, 2021

### §13.1 Sampling from Normal Distribution

**Fact 13.1.** Let  $X_1, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$  population,  $\bar{X}$  be the sample mean and  $S^2$  be the sample variance. Then,

1.  $\bar{X}$  and  $S^2$  are **independent** random variables
2.  $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$
3.  $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$

Recall that if a random vector  $(X, Y)$  has a normal distribution, then

$$X, Y \text{ are independent} \iff \text{Cov}(X, Y) = 0$$

We can actually generalize this idea to linear combinations of normal random variables. Let  $X_1, \dots, X_n$  be random variables such that  $X_i \sim N(\mu_i, \sigma_i^2)$  for each  $i = 1, \dots, n$ . Define the random vector  $\mathbf{U} = (U_1, \dots, U_k)$  and  $\mathbf{V} = (V_1, \dots, V_r)$  where

$$U_i = \sum_{j=1}^n a_{ij} X_j, \quad i = 1, \dots, k; \quad V_r = \sum_{j=1}^n b_{rj} X_j, \quad r = 1, \dots, m$$

and  $a_{ij}, b_{rj}$  are constants. Then,

$$U_i \sim N\left(\sum_{j=1}^n a_{ij} \mu_j, \sum_{j=1}^n a_{ij}^2 \sigma_j^2\right), \quad V_r \sim N\left(\sum_{j=1}^n b_{rj} \mu_j, \sum_{j=1}^n b_{rj}^2 \sigma_j^2\right), \quad \text{Cov}(U_i, V_r) = \sum_{j=1}^n a_{ij} b_{rj} \sigma_j^2$$

and,

$$U_i, V_r \text{ are independent} \iff \text{Cov}(U_i, V_r) = 0$$

And this also implies that

$$\mathbf{U}, \mathbf{V} \text{ are independent random vectors} \iff U_i, V_r \text{ are independent } \forall i, \forall r.$$

**Application:** If  $X_1, \dots, X_n$  is a random sample from  $N(\mu, \sigma^2)$  population,

$$X_j - \bar{X} \text{ and } \bar{X} \text{ are independent}$$

for every  $j = 1, \dots, n$ . From here we conclude that  $S^2$  and  $\bar{X}$  are independent.

### §13.2 The Derived Distributions: Student's $t$ and Snedecor's $F$

**Definition 13.1 (Student's  $t$  distribution)** We say that a random variable  $T$  has **Student's  $t$  distribution** with  $p$  degrees of freedom (and we write  $T \sim t_p$ ) if its pdf is given by

$$f(t) = \frac{\Gamma(p+1)/2}{\Gamma(p/2)} \cdot \frac{1}{(1+t^2/p)^{(p+1)/2}}, \quad -\infty < t < \infty.$$

#### Properties of the $t$ distribution

1.  $t_1 = \text{Cauchy}(0, 1)$
2. the graph of student's  $t$  distribution is bell-shaped and symmetric around 0.
3. We have  $\mathbb{E}T = 0$ ,  $\text{Var}T = \frac{p}{p-2}$  if  $p > 2$ .

Let  $U$  and  $V$  be random variables such that  $U \sim \text{Normal}(0, 1)$  and  $V \sim \chi_p^2$ . Then

$$T := \frac{U}{\sqrt{V/p}} \sim t_p.$$

**Application.** Let  $X_1, \dots, X_n$  be a random sample from  $\text{Normal}(\mu, \sigma^2)$  population. Then

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

**T is used to make inferences about  $\mu$ , when  $\sigma^2$  is unknown.**

**Definition 13.2 (Snedecor's F distribution)** We say that a continuous random variable  $F$  has a Snedecor's F distribution with  $p$  and  $q$  degrees of freedom (and we write  $F \sim F_{p,q}$ ) if its pdf is given by

$$f(x) = \frac{\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \left(\frac{p}{q}\right)^{p/2} \cdot \frac{x^{(p/2)-1}}{[1 + (p/q)x]^{(p+q)/2}}, \quad 0 < x < \infty.$$

Let  $U$  and  $V$  be independent random variables such that  $U \sim \chi_p^2$  and  $V \sim \chi_q^2$ . Then

$$F := \frac{U/p}{V/p} \sim F_{p,q}.$$

**Application.** Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random sample from  $\text{Normal}(\mu_X, \sigma_X^2)$  and  $\mathbf{Y} = (Y_1, \dots, Y_n)$  be a random sample from a  $\text{Normal}(\mu_Y, \sigma_Y^2)$ . Suppose that  $\mathbf{X}$  and  $\mathbf{Y}$  are independent. Then,

$$F := \frac{S_X^2/\sigma^2}{S_Y^2/\sigma_Y^2} \sim F_{m-1, n-1}$$

$F$  is used to make inferences about the ratio of  $\sigma_X^2/\sigma_Y^2$

## §14 November 8, 2021

### §14.1 Order Statistics

We will consider a transformation that takes  $n$  RV's  $X_1, \dots, X_n$  and essentially returns them in a sorted order.

**Definition 14.1** The **order statistics** of random variables  $X_1, \dots, X_n$  are the random variables  $X_{(1)}, \dots, X_{(n)}$ , where

$$\begin{aligned} X_1 &= \min(X_{(1)}, \dots, X_{(n)}) \\ X_2 &= \text{is the second-smallest of } X_1, \dots, X_n \\ &\vdots \\ X_{n-1} &= \text{is the second-largest of } X_1, \dots, X_n \\ X_n &= \max(X_{(1)}, \dots, X_{(n)}). \end{aligned}$$

The sample range is a statistic defined as  $R = X_{(n)} - X_{(1)}$ . The midrange statistic is defined as  $V = (X_{(1)} + X_{(n)})/2$ . The *sample median* is defined by

$$M = \begin{cases} X_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}) & \text{if } n \text{ is even} \end{cases}$$

It is important to note that order statistics  $X_{(1)}, \dots, X_{(n)}$  are random variables, and each  $X_{(i)}$  is a function of the random sample  $X_1, \dots, X_n$ . Even if the original sample is independent, the order statistics are **dependent**!

**Theorem 14.2** — Let  $X_1, \dots, X_n$  be a random sample from discrete distribution with pmf  $f(x)$ . Suppose that the sample space of  $X_1$  is a set  $\{x_1, \dots, x_n\}$  such that  $x_1 < x_2 < \dots$  with CDF  $F(x)$ . Then the CDF of the  $j^{\text{th}}$  order statistic  $X_{(j)}$  is.

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

The pdf of the  $j^{\text{th}}$  order statistic  $X_{(j)}$  is given by

$$f(X_{(j)}(x)) = \frac{n!}{(j-1)!(n-j)!} f(x) [F(x)]^{j-1} [1 - F(x)]^{n-j}$$

Let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics of a random sample,  $X_1, \dots, X_n$ , from continuous population with cdf  $F_X(x)$  and pdf  $f_X(x)$ . Then the joint pdf of  $X_{(i)}$  and  $X_{(j)}$ ,  $1 \leq i < j \leq n$ , is

$$f_{X_{(i)}, X_{(j)}}(u, v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} \times [F_X(u) - F_X(v)]^{j-1-i} [1 - F_X(v)]^{n-j}$$

## §15 November 10, 2021

### §15.1 Convergence Concepts

The notion of letting sample size approach infinity can provide us with useful approximations, since it usually happens that expressions become simplified in the limit.

### §15.2 Convergence in Probability

**Definition 15.1 (Convergence in probability)** A sequence of random variables,  $X_1, X_2, \dots$  converges in probability to a random variable  $X$  if, for every  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0 \quad \text{or equivalently,} \quad \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

The law of large number asserts that as  $n$  grows, the sample mean  $\bar{X}_n$  converges to true mean  $\mu$ . Law of larger number has two versions (weak and strong), the difference in the two lies in what is mean for a sequence of random variables to converge to a number.

**Theorem 15.2 (Weak Law of Large numbers)** — Let  $X_1, X_2, \dots$  be iid random variables with  $\mathbb{E}X_i = \mu$  and  $\text{Var}X_i = \sigma^2 < \infty$ . Define,  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Then, for every  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| < \varepsilon) = 1,$$

that is,  $\bar{X}_n$  converges in probability to  $\mu$ .

We can extend the definition of convergence in probability to functions of random variables. Suppose that  $X_1, X_2, \dots$  converges in probability to a random variable  $X$  or to a constant  $a$  and  $h$  is a continuous function. Then  $h(X_1), h(X_2), \dots$  converges in probability to  $h(X)$ . We can use the above fact to easily prove that since  $S_n^2 \rightarrow \sigma^2 \implies S_n = \sqrt{S_n^2} \rightarrow \sigma$ .

**Properties of convergence in probability:**

1. if  $X_n \xrightarrow{p} X$  and  $X_n \xrightarrow{p} Y$ , then  $X = Y$  asymptotically.
2. if  $X_n \xrightarrow{p} X$  and if  $Y_n \xrightarrow{p} Y$ , then if  $X_n + Y_n \xrightarrow{p} X + Y$ .
3. If  $X_n \xrightarrow{p} X$ ,  $Y_n \xrightarrow{p} Y$ , and  $g(x, y)$  is a continuous function, then  $g(X_n, Y_n) \xrightarrow{p} g(X, Y)$ .

**§15.3 Almost Sure Convergence and Strong Law of Large Numbers**

Almost sure convergence is stronger than convergence in probability. It is similar to point wise convergence of a sequence of functions.

**Definition 15.3 (Almost surely convergence)** A sequence of random variables,  $X_1, X_2, \dots$  converges almost surely to a random variable  $X$  if, for every  $\varepsilon > 0$ ,

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1.$$

Lets try to understand this definition a bit deeper. A random-variable is just a real-valued function defined on a sample space  $S$ . Let  $s \in S$ , then  $X_n(s)$  and  $X(s)$  are all functions defined on  $S$ . The definition is saying that  $X_n$  converges to  $X$  almost surely if the functions  $X_n(s)$  converges to  $X(s)$ .

*Note:* *Almost surely convergence* implies convergence in probability, however converse is not true. We say that the sequence of  $\{\hat{\theta}_n\}_n$  is a *strongly consistent* estimators for the parameter  $\theta$  if  $\hat{\theta}_n \xrightarrow{a.s.} \theta$ .

**Theorem 15.4 (The Strong Law of Large Numbers)** — Let  $\{\bar{X}_n\}_n$  be a sequence of iid random variables. Suppose that  $\mathbb{E}|X_1| \leq \infty$  and  $\mathbb{E}X_1 = \mu$ . Then

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

In other words, sample mean  $\bar{X}_n$  converges to the true mean  $\mu$  pointwise as  $n \rightarrow \infty$ .

The law of large number plays a crucial role in simulations and statistics. Lets say we generate data from a large number of i.i.d samples of an experiment, either using a computer simulation or relation world. If we employ proportion of times an event occurred to approximate the probability of the event, we are *implicitly* applying the Law of Large Numbers.

**§15.4 Convergence in Distribution and Central Limit Theorem**



**Definition 15.5 (Convergence in Distribution)** We say that  $\{X_n\}_{n \geq 1} = \{X_1, X_2, \dots\}$  of random variables *converges in distribution* to a random variable  $X$  if

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x) \xrightarrow{n \rightarrow \infty} \mathbb{P}(X \leq x) = F_X(x) \quad \text{for all } x \in \mathbb{R}$$

such that  $F_X$  is continuous.

We say that sequence  $\hat{\theta}_{n \geq 1}$  of estimators of  $\theta$  is **asymptotically normal** for  $\theta$  if

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} Z \sim N(0, \sigma^2) \quad \text{for some } \sigma^2 > 0.$$

Lets try to connect the ideas from previous sections to above definition. The law of large numbers essentially says that if we have  $X_1, X_2, X_3, \dots$  i.i.d with mean  $\mu$  and variance  $\sigma^2$ ,  $\bar{X}_n \rightarrow \mu$  as  $n \rightarrow \infty$  with probability 1. But what is the distribution of  $\bar{X}_n$  along the way to becoming a constant? This is where the Central Limit Theorem (CLT) comes into play.

**Theorem 15.6 (Central Limit Theorem)** — Let  $\{X_n\}_{n \geq 1} = \{X_1, X_2, \dots\}$  of random variables, assume that  $\mathbb{E}X_i = \mu$  is finite and  $\text{Var}(X_i) = \sigma^2 < \infty$ . Then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} Z \sim N(0, \sigma^2)$$

If  $X_n \xrightarrow{d} X$  and  $X_n \xrightarrow{d} a$ , then

$$X_n + Y_n \xrightarrow{d} X + a \quad \text{and} \quad X_n Y_n \xrightarrow{d} Xa$$

We say that  $\{X_n\}_{n \geq 1}$  is a sequence of asymptotically normal estimators of  $\mu$ .

**Delta Method.** Let  $\{\hat{\theta}_n\}_n$  be a sequence of asymptotically normal estimators of  $\theta$ . Recall that asymptotically normal means that random variables satisfies  $\sqrt{n}(Y_n - \theta) \rightarrow n(0, \sigma)$  in distribution. For a given function  $g$  and specific value of  $\theta$ , suppose that  $g'(\theta)$  exists and is not 0. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \rightarrow n(0, \sigma^2[g'(\theta)]^2) \quad \text{in distribution.}$$

## §16 November 15, 2021

### §16.1 Inference

Our objective is now to use the information in the sample  $X_1, \dots, X_n$  to make inferences about the unknown parameter.

For  $j = 1, \dots, m$ , let  $T_j$  be measurable functions defined on  $\mathbb{R}^n \rightarrow \mathbb{R}$  and not depending on  $\theta$ , and let  $\mathbf{T} = (T_1, \dots, T_m)'$ . Then

$$\mathbf{T}(X_1, \dots, X_n) = (T_1(X_1, \dots, X_n), \dots, T_m(X_1, \dots, X_n))'$$

is called a **m-dimensional statistic**. Any statistic  $\mathbf{T}(\mathbf{X})$  defines a form of data reduction by partitioning the sample space  $\mathcal{X}$ . If we only use the value of the statistic,  $\mathbf{T}(\mathbf{x})$ , rather than the entire sample  $\mathbf{x}$ , the two samples  $\mathbf{x}$  and  $\mathbf{y}$  will be treated equally if  $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$

## §16.2 Sufficiency Principle

**Definition 16.1** Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random sample with a joint pdf or pmf denoted by  $f(\mathbf{x}|\theta)$ . Let  $T = T(\mathbf{x})$  be a statistic based on this sample with pdf (or pmf) denoted by  $f_T(t|\theta)$ . We say that  $T$  is a **sufficient statistic** for  $\theta$  if for any  $t$  such that  $f_T(t|\theta) > 0$ , the conditional pdf (or pmf) of  $\mathbf{X}$  given  $T = t$  does not depend on  $\theta$ .

*Remark.* The joint pdf (or pmf) of  $(\mathbf{X}, T)$  is

$$f_{\mathbf{X},T}(\mathbf{x}, t|\theta) = \begin{cases} f(\mathbf{x}|\theta) & \text{if } t = T(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

**Theorem 16.2 (Factorization Theorem)** — Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random sample with a joint pdf or pmf denoted by  $f(\mathbf{x}|\theta)$ . Let

$$\mathbf{T} = \mathbf{T}(X_1, \dots, X_n) = (T_1(X_1, \dots, X_n), \dots, T_k(X_1, \dots, X_n))'$$

be a  $k$ -dimensional statistic. Then  $\mathbf{T}$  is a sufficient statistic for  $\theta$  if and only if

$$f(\mathbf{x}|\theta) = g(\mathbf{T}|\theta) \cdot h(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{X}.$$

## §17 November 17, 2021

### §17.1 Factorization Theorem Continued

**Example 17.1** — Let  $X_1, \dots, X_n$  be independent random variables such that

$$X_k \sim \text{Unif}(k(\theta - 1), k(\theta + 1)).$$

Show that  $\left( \min_{1 \leq k \leq n} \frac{x_k}{k}, \max_{1 \leq k \leq n} \frac{x_k}{k} \right)$  is a 2-dimensional sufficient statistic.

**Theorem 17.2** — Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random sample from a pdf(or pmf) which belongs to the following exponential family

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left\{\sum_{i=1}^k w_i(\theta)t_i(\mathbf{x})\right\}$$

with  $\theta = (\theta_1, \dots, \theta_d)$ . Suppose that  $d \leq k$ . Define

$$T_i = T_i(\mathbf{X}) = \sum_{j=1}^n T_i(X_j) \quad \forall i = 1, \dots, k.$$

Then  $\mathbf{T} = (T_1, \dots, T_k)$  is a sufficient statistic for  $\theta$ .

*Proof.* The joint pdf of  $\mathbf{X}$  is:

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{j=1}^n f_{x_j}(x_j|\theta) = \prod_{j=1}^n \left[ h(x_j)c(\theta) \left\{ \sum_{i=1}^k w_i(\theta)t_i(x_j) \right\} \right] \\ &= \prod_{j=1}^n h(x_j)[c(\theta)]^n \exp \left\{ \sum_{i=1}^k w_i(\theta)t_i(x_j) \right\} \end{aligned}$$

Clearly,

$$\underbrace{\prod_{j=1}^n h(x_j)}_{h'(\mathbf{x})} \underbrace{[c(\theta)]^n \exp \left\{ \sum_{i=1}^k w_i(\theta)t_i(x_j) \right\}}_{g(T_1(\mathbf{x}), \dots, T_k(\mathbf{x})|\theta)}$$

Thus, by the Factorization theorem  $(T_1(\mathbf{x}), \dots, T_k(\mathbf{x})|\theta)$  is a sufficient statistic for  $\theta$ .  $\square$

**Remark:** If  $\mathbf{T}$  is a sufficient statistic for  $\theta$ , then any one-to-one transformation of  $\mathbf{S} = \pi(\mathbf{T})$  is also a sufficient statistic for  $\theta$ ,  $\forall \pi$  one-to-one maps. **Therefore, a sufficient statistic is not unique.**

**Definition 17.3** A sufficient statistic is  $\mathbf{T} = (T_1, \dots, T_k)$  is called a **minimal sufficient statistic** for  $\theta$  if, for any other sufficient statistic  $\mathbf{T}^* = (T_1^*, \dots, T_k^*)$ , there exists a function  $\phi$  such that  $\mathbf{T} = \phi(\mathbf{T}^*)$ . This is equivalent to saying that

$$T^*(\mathbf{x}) = T^*(\mathbf{y}) \implies T(\mathbf{x}) = T(\mathbf{y})$$

**Note:** A minimal sufficient statistic may **NOT** be unique.

**Theorem 17.4 (Lehman and Scheffe)** — Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random sample from a pdf(or pmf) which belongs to the following exponential family  $f_X(x|\theta)$ . Let  $T = T(\mathbf{X})$  be a statistic which satisfies the following condition

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} \text{ does not depend on } \theta \iff T(\mathbf{x}) = T(\mathbf{y})$$

Then  $T$  is a minimal sufficient statistic.

*Proof.* Proof is given in the text book.  $\square$

## §18 November 22, 2021

### §18.1 The Sufficiency Principle (Continued)

So far, we have covered sufficient statistics which in a sense contain all the information about  $\theta$  that is available in the sample. Next we look at a statistic which is quite the opposite.

**Definition 18.1 (Ancillary Statistic)** A statistic  $S(\mathbf{X})$  whose distribution does not depend on the parameter  $\theta$  is called an **ancillary** statistic.

An ancillary statistic contains no information about  $\theta$ ! An ancillary statistic has a fixed and known distribution that is unrelated to  $\theta$ .

**Example 18.2** (Location and Scale family ancillary statistic) — Let  $X_1, \dots, X_n$  be iid observations from a location parameter family with pdf  $g(x|\mu) = f(x - \mu)$  where  $f$  is a standard pdf. Then we will show that

$$\begin{cases} R = X_{(n)} - N_{(1)} & \text{is ancillary statistic for } \mu. \\ S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 & \text{is ancillary statistic for } \mu. \end{cases}$$

In addition to that if we have random sample  $\mathbf{X} = (X_1, \dots, X_n)$  from a *scale family* with pdf  $g(x|\theta) = \frac{1}{\sigma} f(x/\sigma)$ , where  $f(z)$  is the standard pdf of the family. Then

$$T(\mathbf{X}) = \left( \frac{X_1}{X_n}, \dots, \frac{X_{n-1}}{X_n} \right) \text{ is an ancillary statistic for } \sigma$$

In particular  $\bar{X}/X_n$  is an ancillary statistic for  $\sigma$ .

*Proof.* □

**Definition 18.3 (Complete statistic)** Let  $f(t|\theta)$  be a family of pdf or omfs for a statistic  $T(\mathbf{X})$ . The family of probability distributions is called **complete** if  $\mathbb{E}_\theta(T) = 0$  for all  $\theta$  implies that  $\mathbb{P}_\theta(g(T) = 0) = 1$  for all  $\theta$ . Equivalently,  $T(\mathbf{X})$  is called a **complete statistic**.

In other words, let  $T$  be the statistic whose range is  $\mathcal{T}$ , it is called complete if

$$\mathbb{E}g(T) = 0 \text{ for all } \theta \implies g(t) = 0 \text{ for all } t \in \mathcal{T}$$

**Theorem 18.4 (Basu's Theorem)** — If  $T(\mathbf{X})$  is a complete ad minimal sufficient statistic, the  $T(\mathbf{X})$  is independent of every ancillary statistic.

*Proof.* The proof for the discrete case is given on page 287 of the textbook, review it! □

**Theorem 18.5 (Complete Statistics in the exponential families)** — Let  $\mathbf{X} = (X_1, \dots, X_n)$  be iid observations from an exponential family with pdf or pmf of the form

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^k w(\theta_j)t_j(x),\right)$$

where  $\theta = (\theta_1, \dots, \theta_k)$ . Then the statistic

$$T(\mathbf{X}) = \left( \sum_{i=1}^n t_1(X_i), \sum_{i=1}^n t_2(X_i), \dots, \sum_{i=1}^n t_k(X_i), \right)$$

is complete as long as the parameter space  $\Theta$  contains an open set in  $\mathbb{R}^k$ .

*Proof.* Proof is omitted from this course. □

## §18.2 Likelihood Principle

**Definition 18.6 (Likelihood function)** Let  $f(\mathbf{x}|\theta)$  denote the joint pdf or pmf of the sample  $\mathbf{X} = (X_1, \dots, X_n)$ . Then, given that  $\mathbf{X} = \mathbf{x}$  is observed, the function of  $\theta$  defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the **likelihood function**.

**Likelihood Principle:** If  $\mathbf{x}$  and  $\mathbf{y}$  are two sample points such that  $L(\theta|\mathbf{x}) \propto L(\theta|\mathbf{y})$ , i.e.,  $\exists C(\mathbf{x}, \mathbf{y})$  constant such that

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y})L(\theta|\mathbf{y}) \quad \text{for all } \theta,$$

then the conclusions drawn from  $\mathbf{x}$  and  $\mathbf{y}$  should be identical. We define an **experiment**  $E$  to be a triple  $(\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ , where  $\mathbf{X}$  is a random vector with pmf  $f(\mathbf{x}|\theta)$  for some  $\theta \in \Theta$ . An experimenter knowing what experiment  $E$  was performed and having observed a particular  $\mathbf{X} = \mathbf{x}$ , will make some inference or draw some conclusion about  $\theta$ . We denote this conclusion by  $\text{Ev}(E, \mathbf{x})$ , which stands for *evidence about  $\theta$  arising from  $E$  and  $\mathbf{x}$* .

**Formal sufficiency Principle:** Consider experiment  $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$  and suppose that  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ . If  $\mathbf{x}$  and  $\mathbf{y}$  are two samples such that  $T(\mathbf{x}) = T(\mathbf{y})$  then

$$\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y})$$

**Formal Likelihood Principle:** Suppose that we have two experiments,  $E_1 = (\mathbf{X}_1, \theta, \{f(\mathbf{x}_1|\theta)\})$  and  $E_2 = (\mathbf{X}_2, \theta, \{f(\mathbf{x}_2|\theta)\})$ , where we have the unknown parameter  $\theta$  is the same for both experiments. If  $\mathbf{x}$  is a sample from  $E_1$  and  $\mathbf{y}$  is a sample from  $E_2$  such that

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y})L(\theta|\mathbf{y}) \quad \text{for all } \theta,$$

then,

$$\text{Ev}(E_1, \mathbf{x}) = \text{Ev}(E_2, \mathbf{y})$$

## §19 November 24, 2021

### §19.1 Equivariance Principle

**Equivariance Principle:** If  $\mathbf{Y} = g(\mathbf{X})$  is a change of measurement scale such that the model for  $\mathbf{Y}$  has the same formal structure as the model for  $\mathbf{X}$ , then inference procedure should be both measurement equivariant and formally equivariant.

**Definition 19.1 (Group transformation)** A set  $\mathcal{G}$  of functions,  $\{g(\mathbf{x} : g \in \mathcal{G})\}$ , of the form  $g : \mathcal{X} \rightarrow \mathcal{X}$  is called a **group transformation** of  $\mathcal{X}$  if

- (i) (Inverse)  $\forall g \in \mathcal{G}, \exists g^{-1} \in \mathcal{G}$  such that  $g \circ g^{-1} = e$ , where  $e : \mathcal{X} \rightarrow \mathcal{X}$  is the identity  $e(\mathbf{x}) = \mathbf{x}$ .
- (ii) (Composition) For every  $g, g' \in \mathcal{G}, g \circ g' \in \mathcal{G}$
- (iii) (Identity) The identity,  $e(\mathbf{x})$  is an element of  $\mathcal{G}$ .

**Definition 19.2 (Invariant)** Let  $\mathcal{F} = \{f(\mathbf{x}|\theta) : \theta \in \Theta\}$  be a set of pdf or pmfs for  $\mathbf{X}$ , and let  $\mathcal{G}$  be a group of transformations of the sample space  $\mathcal{X}$ . Then  $\mathcal{F}$  is **invariant under the group  $G$**  if for every  $\theta \in \Theta$  and  $g \in \mathcal{G}$  there exists a unique  $\theta' \in \Theta$  such that  $\mathbf{Y} = g(\mathbf{X})$  has the distribution  $f(\mathbf{y}|\theta')$  if  $\mathbf{X}$  has the distribution  $f(\mathbf{x}|\theta)$ .

The equivariance principle essentially says that inference based on  $\mathbf{X}$  should be the same as inference based on  $\mathbf{Y}$ .

**Example 19.3** (Location family is invariant) —

**Example 19.4** (Scale family is invariant) —

## §19.2 Methods for finding estimators

A statistic is a function of the random sample, and we generally use such functions to approximate the value of the unknown parameter  $\theta$ , which is underlying the parameter for the underlying distribution of the sample) called **estimator**. Once we have observed a particular realization of  $\mathbf{X}$ , (the sample  $\mathbf{x}$ ), we refer to the value  $T(\mathbf{x})$  of the estimator for that particular realization as **estimate for  $\theta$** .

We will study **3 methods for constructing estimator**: method of moments, maximum likelihood estimation, and Bayes method.

## §19.3 Methods of Moments

Let  $X_1, \dots, X_n$  be a sample from a population with pdf or pmf  $f(x|\theta_1, \dots, \theta_k)$ . **Method of moment estimators** are found by equating the first  $k$  sample moments to corresponding  $k$  population moments.

$$\begin{aligned} m_1 &= \frac{1}{n} \sum_{i=1}^n X_i^1, & \mu_1(\theta) &= \mathbb{E}X^1 = \int_{-\infty}^{\infty} x f(x|\theta) dx \\ m_2 &= \frac{1}{n} \sum_{i=1}^n X_i^2, & \mu_2(\theta) &= \mathbb{E}X^2 = \int_{-\infty}^{\infty} x^2 f(x|\theta) dx \\ &\vdots & & \\ m_k &= \frac{1}{n} \sum_{i=1}^n X_i^k, & \mu_k(\theta) &= \mathbb{E}X^k = \int_{-\infty}^{\infty} x^k f(x|\theta) dx \end{aligned}$$

The method of moment estimator  $(\hat{\theta}_1, \dots, \hat{\theta}_k)$  of  $(\theta_1, \dots, \theta_k)$  is the solution of the full system of  $k$  equations:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n X_i^1 &= \mu_1(\theta) \\ \frac{1}{n} \sum_{i=1}^n X_i^2 &= \mu_2(\theta) \\ &\vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^k &= \mu_k(\theta) \end{aligned}$$

**Remark:**  $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2\bar{X} \sum_{i=1}^n X_i = \sum_{i=1}^n X_i^2 - n\bar{X}^2$

$$\implies \sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2$$

## §19.4 Maximum Likelihood Estimators

The method of maximum likelihood is by far the most popular technique for deriving estimators. For each observed sample  $\mathbf{x} = (x_1, \dots, x_n)$ , we define the **maximum likelihood estimate** as the point  $\hat{\theta} = \hat{\theta}(\mathbf{x})$  where the function  $\theta \mapsto L(\theta|\mathbf{x})$  attains its maximum, i.e.:

$$\max_{\theta \in \Theta} L(\theta|\mathbf{x}) = L(\hat{\theta}|\mathbf{x})$$

For each sample point  $\mathbf{x}$ , and  $\hat{\theta}(\mathbf{x})$  is the parameter value at which  $L(\theta|\mathbf{x})$  attains its maximum as a function of  $\theta$ , with  $\mathbf{x}$  held fixed. A *maximum likelihood estimator (MLE)* of the parameter  $\theta$  based on sample  $\mathbf{X}$  is  $\hat{\theta}(\mathbf{X})$ .

To find the MLE, in most cases we will be solving the equation

$$\frac{d}{d\theta} L(\theta|\mathbf{x}) = 0$$

and checking that

$$\frac{d^2}{d\theta^2} L(\theta|\mathbf{x}) \big|_{\theta=\hat{\theta}} < 0$$

## §20 November 29, 2021

### §20.1 Invariance property of MLE

Suppose that the distribution is indexed by the parameter  $\theta$ , if we are interested in estimating some function of  $\theta$ , say  $\tau(\theta)$ . If we let  $\eta = \tau(\theta)$ , then the inverse function  $\theta = \tau^{-1}(\eta)$  is well-defined, and we can express the likelihood function as a function of  $\eta$

$$L^*(\eta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\mathbf{x})$$

and,

$$\sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\eta} L(\tau^{-1}(\eta)|\mathbf{x}) = \sup_{\eta} L(\theta|\mathbf{x})$$

The **induced likelihood** of  $\eta$  is defined by

$$\sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\theta: \tau(\theta)=\eta} L(\theta|\mathbf{x})$$

The value of  $\hat{\eta}$  which maximized  $L^*(\eta|\mathbf{x})$  is called the MLE of  $\eta$ :

$$\sup_{\eta} L^*(\eta|\mathbf{x}) = L^*(\hat{\eta}|\mathbf{x})$$

**Theorem 20.1** (Invariance property of MLEs) — If  $\hat{\theta}$  is the MLE of  $\theta$ , then for any function  $\tau(\theta)$ , the MLE of  $\tau(\theta)$  is  $\tau(\hat{\theta})$ .

*Proof.* Let  $\hat{\eta}$  denote the value that maximizes  $L^*(\eta|\mathbf{x})$ . Our objective is to show that  $L^*(\eta|\mathbf{x}) = L^*(\tau(\hat{\theta})|\mathbf{x})$ . Since,

$$\begin{aligned} L^*(\hat{\eta}|\mathbf{x}) &= \sup_{\eta} \sup_{\theta: \tau(\theta)=\eta} L(\theta|\mathbf{x}) \\ &= \sup_{\theta} L(\theta|\mathbf{x}) \\ &= L(\hat{\theta}|\mathbf{x}) \end{aligned}$$

□

We can see that using this theorem that MLE of  $\theta^2$  is  $\bar{X}^2$ , and for a more complicated function such as  $\sqrt{p(1-p)}$ , where  $p$  is the binomial probability, the MLE is  $\sqrt{\hat{p}(1-\hat{p})}$ .

## §20.2 Bayes Estimators

Denote the prior distribution  $\pi(\theta)$  and the sampling distribution by  $f(\mathbf{x}|\theta)$ , then the posterior distribution is given by

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \quad (\text{Bayes' Rule!})$$

where  $m(\mathbf{x})$  is the marginal distribution of  $\mathbf{X}$ , that is

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$$

**Definition 20.2** Let  $\mathcal{F}$  be a class of distributions for  $X$ . A class  $\Pi$  of prior distributions is a **conjugate family** for  $\mathcal{F}$  if the posterior distribution is in the class  $\Pi$  for all  $f \in \mathcal{F}$ , i.e.,

$$\pi(\theta|\mathbf{x}) \quad \text{for all } \pi \in \Pi, f \in \mathcal{F}$$

The **Bayes estimator** of  $\theta$  is defined as the expected value of  $\theta$  under the posterior distribution, which is the weighted average value of  $\theta$  given the new evidence we have after observing the sample;

$$\hat{\theta}_B(\mathbf{x}) = \int \theta \pi(\theta|\mathbf{x}) d\theta$$

**Example 20.3** (Conjugate families) — We showed that

$$\pi(\theta|\mathbf{x}) = \text{Beta} \left( \alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i \right)$$

thus, we can clearly see that the beta family is a conjugate for the Bernoulli family. In addition to that we showed the **normal family** is conjugate to itself since

$$\pi(\theta|\mathbf{x}) = \text{Normal} \left( \frac{n\tau^2\bar{x} + \sigma^2\mu}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} \right)$$



## §21 December 1, 2021

### §21.1 Methods of evaluating estimators

The general topic of evaluating statistical procedures is part of a branch of statistics called decision theory. In this section we look at some basic criteria for evaluating estimators.

**Definition 21.1 (Mean square error (MSE))** The **mean square error** of an estimator  $W$  of a parameter  $\theta$  is the function of  $\theta$  defined by  $\mathbb{E}_\theta(W - \theta)^2$ .

**Definition 21.2 (Bias)** The bias of an estimator  $T$  for a parameter  $\theta$  is defined by

$$\text{Bias}_\theta(T) = \mathbb{E}_\theta T - \theta$$

**Note:**  $MSE$  incorporates two components, one measuring the variability and the other measuring bias.

$$MSE_\theta = \text{Var}_\theta(T) + (\text{Bias}_\theta(T))^2$$

**Example 21.3** — Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ . The statistics  $\bar{X}$  and  $S^2$  are both biased estimators since

$$\mathbb{E}\bar{X} = \mu, \mathbb{E}S^2 = \sigma^2, \quad \text{for all } \mu \text{ and } \sigma.$$

The MSEs of the estimators are given by

$$\mathbb{E}(\bar{X} - \mu)^2 = \text{Var}\bar{X} = \frac{\sigma}{n}.$$

$$\mathbb{E}(S^2 - \sigma^2)^2 = \text{Var}S^2$$

### §21.2 Best unbiased estimators (UMVUE)

If we compare estimators based on MSE, there is no single "best MSE" estimator. The reason is that the class of all estimators is too large of a class. Therefore by placing certain restrictions on our estimators, we can limit the class of estimators.

**Definition 21.4 (Best unbiased estimator)** An estimator  $W^*$  is the **best unbiased estimator** of  $\tau(\theta)$  if it satisfies  $\mathbb{E}_\theta W^* = \tau(\theta)$  for all  $\theta$  and, for any other estimator  $W$  with  $\mathbb{E}_\theta W = \tau(\theta)$ , we have  $\text{Var}_\theta W^* \leq \text{Var}_\theta W$  for all  $\theta$ .  $W^*$  is also called a **uniform minimum variance unbiased estimator (UMVUE)**

**Theorem 21.5 (Cramèr-Rao Lower Bound)** — Let  $X_1, \dots, X_n$  be a random variables with joint pdf  $f(\mathbf{x}|\theta)$ . Let  $W = W(\mathbf{X})$  be a finite variance estimator satisfying the following conditio:

$$\frac{d}{d\theta} \int W(\mathbf{x})f(\mathbf{x}|\theta)d\mathbf{x} = \int W(\mathbf{x})\frac{d}{d\theta}f(\mathbf{x}|\theta)d\mathbf{x}.$$

Then

$$\text{Var}_{\theta}W \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_{\theta}W\right)^2}{I_n(\theta)} = \text{CR-Lower-Bound}$$

Where

$$I_n(\theta) = \mathbb{E}_{\theta} \left( \frac{d}{d\theta} \log f(\mathbf{X}|\theta) \right)^2 = \text{Fisher information}$$

**Note:**

- If  $W$  is an unbiased estimator of  $\theta$  then  $\frac{d}{d\theta}\mathbb{E}_{\theta}W = 1$ , and thus the CR-Lower-Bound becomes  $\frac{1}{I_n(\theta)}$ .
- Generally the condition in the theorem does not hold if the support of  $f(\mathbf{x}|\theta)$  depends on  $\theta$ .
- If  $X_1, \dots, X_n$  are i.i.d with pdf  $f(x|\theta)$ , then

$$I_n(\theta) = nI_1(\theta)$$

- $X_1, \dots, X_n$  are i.i.d with pdf  $f(x|\theta)$  belonging to an exponential family, then the fist condition in the theorem is satisfied and

$$I_1(\theta) = -\mathbb{E}_{\theta} \left( \frac{d^2}{d\theta^2} \log f(\mathbf{X}|\theta) \right)$$

## §22 December 6, 2021

### §22.1 methods of evaluating estimators (continued)

**Corollary 22.1 (Attainment of CR-Lower-Bound)**

If  $X_1, \dots, X_n$  is a random sample from  $f(x|\theta)$  and  $W$  is an bisased estimator of  $\tau(\theta)$  such that condition (1) of Cramèr-Rao theorem is satisfied. Then  $W$  attains the Cramèr-Rao lower bound if and only if there exists a fuction  $a(\theta)$  such that

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{d}{d\theta} \log f(\mathbf{x}|\theta)$$

### §22.2 Sufficiency and Unbiasedness

**Theorem 22.2 (Rao-Blackwell)** — Let  $W$  be any unbiased estimator of  $\tau(\theta)$ , and let  $T$  be a sufficient statistic for  $\theta$ . Define  $\phi(T) = \mathbb{E}(W|T)$ . Then  $\mathbb{E}_\theta \phi(T) = \tau(\theta)$  and  $\text{Var}_\theta \phi(T) \leq \text{Var}_\theta W$  for all  $\theta$ . That is  $\phi(T) = \mathbb{E}(W|T)$  is a uniformly better unbiased estimator of  $\tau(\theta)$ .

**Theorem 22.3** — If  $W$  is the best unbiased estimator of  $\tau(\theta)$ , then  $W$  is unique.

## §23 December 8