

# Statistical Methods for Computational Biology

Sayan Mukherjee



# Contents

<b>Statistical Methods for Computational Biology</b>	1
Lecture 1. Course preliminaries and overview	1
Lecture 2. Probability and statistics overview	3
2.1. Discrete random variables and distributions	3
Lecture 3. Hypothesis testing	25



# Statistical Methods for Computational Biology

Sayan Mukherjee

## LECTURE 1

### Course preliminaries and overview

- Course summary

The use of statistical methods and tools from applied probability to address problems in computational molecular biology. Biological problems in sequence analysis, structure prediction, gene expression analysis, phylogenetic trees, and statistical genetics will be addressed. The following statistical topics and techniques will be used to address the biological problems: classical hypothesis testing, Bayesian hypothesis testing, Multiple hypothesis testing, extremal statistics, Markov chains, continuous Markov processes, Expectation Maximization and imputation, classification methods, and clustering methods. Along the way we'll learn about gambling, card shuffling, and coin tossing.

- Grading

There will be four problem sets that will account for 40% of the grade, a midterm exam that will account for 20% of the grade, a final exam for 40% of the grade (students that have an A after the midterm, both exam and homeworks, will have an option to complete a final project in lieu of the final exam)

---

<sup>1</sup>Institute of Statistics and Decision Sciences (ISDS) and Institute for Genome Sciences and Policy (IGSP), Duke University, Durham, 27708.

**E-mail address:** [sayan@stat.duke.edu](mailto:sayan@stat.duke.edu).

Dec 6, 2005

These lecture notes borrow heavily from many sources.



## LECTURE 2

### Probability and statistics overview

Throughout this course we will quantify, assess, explain noise and randomness in (molecular) biological systems.

Two disciplines will play a prominent role:

- Probability: The word probability derives from the Latin probare (to prove, or to test). Informally, probable is one of several words applied to uncertain events or knowledge, being more or less interchangeable with likely, risky, hazardous, uncertain, and doubtful, depending on the context. Chance, odds, and bet are other words expressing similar notions. As with the theory of mechanics which assigns precise definitions to such everyday terms as work and force, so the theory of probability attempts to quantify the notion of probable.
- Statistics:
  - (1) The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.
  - (2) Numerical data.
  - (3) From German Statistik – political science, from New Latin statisticus – of state affairs, from Italian statista – person skilled in statecraft, from stato, from Latin status – position, form of government.
  - (4) Inverse probability

#### 2.1. Discrete random variables and distributions

**Apology.** *This is not a math course so we will avoid formal mathematical definitions whenever possible.*

##### 2.1.1. Definitions

There exists a *space of elementary events* or *outcomes* of an experiment. The events are discrete (countable) and disjoint and we call this set  $\Omega$ .

**Examples.** *Coin tossing: My experiment is tossing (or spinning) a coin. There are two outcomes “heads” and “tails”. So  $\Omega = \{\text{heads, tails}\}$  and  $|\Omega| = 2$ .*

*Dice rolling: My experiment is rolling a six sided die. So  $\Omega = \{1, 2, 3, 4, 5, 6\}$  and  $|\Omega| = 6$ .*

*Waiting for tail: My experiment is to continually toss a coin until a tail shows up. If I designate heads as  $h$  and tails as  $t$  in a toss then an elementary outcome of this experiment is a sequence of the form  $(hhhhh\dots ht)$ . There are an infinite number of such sequences so I will not write  $\Omega$  and we can state that  $|\Omega| = \infty$ .*

The space of elementary events is  $\Omega$  and the elements of the space  $\Omega$  are denoted as  $\omega$ . Any subset  $A \subseteq \Omega$  is also an event (the event  $A$  occurs if any of the elementary outcomes  $\omega \in A$  occur).

The sum of two events  $A$  and  $B$  is the event  $A \cup B$  consist of elementary outcomes which belong to at least one of the events  $A$  and  $B$ . The product of two events  $A \cap B$  consists of all events belonging to both  $A$  and  $B$ . The space  $\Omega$  is the certain event. The empty set  $\emptyset$  is the impossible event.  $\bar{A} = \Omega - A$  is the complementary event of  $A$ . Two events are mutually exclusive if  $A \cap B = \emptyset$ .

**Examples.** *For a single coin toss coins  $A = t$  and  $B = h$  are complementary and mutually exclusive,  $\Omega - A = t$  and  $t \cap h = \emptyset$ .*

*For a single dice role two possible events are  $A = 1$  and  $B = 6$  so  $A \cup B$  designates either a 1 or 6 is rolled.*

*Consider rolling a die twice the space of elementary events can be designated  $(i, j)$  where  $i, j = 1, \dots, 6$ . The events  $A = \{i + j \leq 3\}$  and  $B = \{j = 6\}$  are mutually exclusive. The product of events  $A$  and  $C = \{j \text{ is even}\}$  is the event  $(1, 2)$ .*

A probability distribution assigns a nonnegative real number to each event  $A \subseteq \Omega$ . A probability distribution is a function  $\Pr$  defined on  $\Omega$  such that  $\Pr : A \rightarrow \mathbb{R}^+$  and

$$\begin{aligned} 1 &= \sum_{\omega \in \Omega} \Pr(\omega) \\ \Pr(A) &= \sum_{\omega \in A} \Pr(\omega). \end{aligned}$$

**Examples.** *Symmetric die, one toss:  $\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = \frac{1}{6}$ .*

*Symmetric coin, one toss:  $\Pr(h) = \Pr(t) = \frac{1}{2}$ .*

*Waiting for head, symmetric coin:  $\Pr(h) = \frac{1}{2}$ ,  $\Pr(th) = \frac{1}{2}^2$ ,  $\Pr(tth) = \frac{1}{2}^3$ , ...  $\sum_{\omega \in \Omega} \Pr(\omega) = \sum_{n=1}^{\infty} 2^{-n} = 1$  so we define a probability distribution on  $\Omega$ . The probability the experiment stops at an even step  $(\cup(th) \cup (tth) \cup \dots)$  is the sum of the corresponding probabilities  $\sum_{n=1}^{\infty} 2^{-2n} = \frac{1}{4} \times \frac{4}{3} = \frac{1}{3}$ .*

Properties of probabilities

- (1)  $\Pr(\emptyset) = 0$  and  $\Pr(\Omega) = 1$
- (2)  $\Pr(A+B) = \sum_{\omega \in A \cup B} \Pr(\omega) = \sum_{\omega \in A} \Pr(\omega) + \sum_{\omega \in B} \Pr(\omega) - \sum_{\omega \in A \cap B} \Pr(\omega) = \Pr(A) + \Pr(B) - \Pr(AB)$
- (3)  $\Pr(\bar{A}) = 1 - \Pr(A)$
- (4) if  $A$  and  $B$  are disjoint events

$$\Pr(A + B) = \Pr(A) + \Pr(B)$$

- (5) Additivity of probability given a set of disjoint events  $A_1, A_2, \dots$  : if  $A_i A_j = \emptyset$  for  $i \neq j$

$$\Pr\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \Pr(A_k)$$

- (6) Subadditivity of probability given a set of events  $A_1, A_2, \dots$  : if

$$\Pr\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \Pr(A_k)$$

Property (6) will be revisited in Multiple Hypothesis Testing and called the Bonferroni correction.

By convention random variables are written in uppercase symbols  $X, Y, Z$  and the observed values in lowercase  $x, y, z$ . Thus the probability distribution can be interpreted as

$$\Pr(y) = \Pr(Y = y).$$

**Definition.** *Distribution function:* The distribution function  $F(y)$  of a random variable  $Y$  is the probability  $Y \leq y$

$$F(Y) = \sum_{y' \leq y} \Pr(y').$$

**Example.**  $Y = \{1, 2, 3\}$  and  $\theta \neq 0$

$$\Pr(y) = \frac{\theta^{2y}}{\theta^2 + \theta^4 + \theta^6}.$$

$$F(1) = \frac{\theta^2}{\theta^2 + \theta^4 + \theta^6}$$

$$F(2) = \frac{\theta^2 + \theta^4}{\theta^2 + \theta^4 + \theta^6}$$

$$F(3) = 1.$$

**Definition.** *Independence:* Two events  $A$  and  $B$  are independent if

$$\Pr(A \cap B) = \Pr(A) \Pr(B).$$

In everyday words this means that the occurrence of event  $A$  does not effect the occurrence of event  $B$ .

**Definition.** *Independence:* Events  $A_1, A_2, \dots, A_n$  and  $B$  are independent if for any  $1 \leq i_1 < i_2 < \dots < i_r < n$ ,  $r = 2, 3, \dots, n$

$$\Pr\left(\bigcap_{k=1}^r A_{i_k}\right) = \prod_{k=1}^r \Pr(A_{i_k}).$$

**Definition.** *Conditional probabilities:* The probability of the event  $A$  given the event  $B$  (assuming  $\Pr(B) > 0$ ) is the conditional probability of  $A$

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)}.$$

**Example.** We toss two fair coins. Let event  $A$  be the event that the first toss results in heads, and event  $B$  the event that tails shows up on the second toss

$$\Pr(AB) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = \Pr(A) \Pr(B).$$

We have a die that is tetrahedron (*Dungeons and Dragons* for example) with three faces painted red, blue, and green respectively, and the fourth in all three colors. We roll the die. Event  $R$  is that the bottom face is red, event  $B$  that it is blue, and event  $G$  it is green.  $\Pr(R) = \Pr(B) = \Pr(G) = \frac{2}{4} = \frac{1}{2}$ .  $\Pr(RB) = \Pr(GB) = \Pr(GR) = \dots = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ . However,

$$\Pr(RGB) = \frac{1}{4} \neq \Pr(R) \Pr(G) \Pr(B) = \frac{1}{8}.$$

**Definition.** Given two integers  $n, i$  with  $i \leq n$  the number of ways  $i$  integers can be selected from  $n$  integers irrespective of order is

$$\binom{n}{i} = \frac{n!}{(n-i)! i!}.$$

This is often stated as  $n$  choose  $i$ .

### 2.1.2. My favorite discrete distributions

- (1) Bernoulli trials: A single trial (coin flip) with two outcomes  $\Omega = \{\text{failure, success}\}$ . The probability of success is  $p$  and therefore failure is  $1 - p$ .

If we assign to a Bernoulli random variable  $Y$  the number of successes in a trial the  $Y = 0$  with probability  $1 - p$  and  $Y = 1$  with  $p$  and

$$\Pr(y) = p^y (1 - p)^{1-y}, \quad y = 0, 1.$$

If we assign to a Bernoulli random variable  $S$  the value  $-1$  if the trial results in failure and  $+1$  if it results in success then

$$\Pr(s) = p^{(1+s)/2} (1 - p)^{(1-s)/2}, \quad s = -1, 1.$$

- (2) Binomial distribution: A binomial random variable is the number of successes of  $n$  independent and identical Bernoulli trials. Identical Bernoulli trials means that the distribution function is identical for each trial or that each trial shares the same probability of success  $p$ . The two variables  $n$  and  $p$  are the index and parameter for the distribution respectively. The probability distribution of  $Y$  is

$$\Pr(y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 1, 2, \dots, n.$$

**Example.** We are given two small DNA sequences of length  $n = 25$  and  $y = 11$  matches.

↓	↓		↓		↓	↓	↓	↓			↓	↓	↓											
g	g	a	g	a	c	t	g	t	a	g	a	c	g	c	t	a	a	t	g	c	t	a	t	a
g	a	a	c	g	c	c	c	t	a	g	c	c	g	g	a	g	c	c	c	t	t	a	t	c

Assuming equal probabilities of  $a, c, t, g$  at each site and independence we know that the chance of a match (success) is  $p = \Pr(aa) + \Pr(cc) + \Pr(gg) + \Pr(tt) = \frac{1}{4}$ . We can now use the binomial distribution to ask how likely would there be 11 matches in 26 nucleotides.

*Is i.i.d. Bernoulli a good idea for nucleotides ?*

- (3) Uniform distribution: The random variable  $Y$  takes the possible values  $\Omega = \{a, a+, \dots, a + b - 1\}$  or  $U[a, b]$  where  $a, b \in \mathbb{N}$  and

$$\Pr(y) = b^{-1} \quad \text{for } y = a, a + 1, \dots, a + b - 1.$$

- (4) Geometric distribution: We run a sequence of i.i.d. Bernoulli trials each with success  $p$ . The random variable of interest the number of trials  $Y$  before (but not including) the first failure. Again  $Y = \{0, 1, 2, \dots\}$ . If  $Y = y$  then there must have been  $y$  successes followed by one failure so the distribution is

$$\Pr(y) = (1 - p)p^y, \quad y = 0, 1, 2, \dots$$

and the distribution function is

$$F(y) = \Pr(Y \leq y) = 1 - p^{y+1}, \quad y = 0, 1, 2, \dots$$

The length of (*sssssssssf*) of  $Y$  is the length of a success run. This type of random variable will be used in computing probabilities random of contiguous matches in the comparison of two sequences.

- (5) Poisson Distribution: A random variable  $Y$  has a Poisson distribution (with parameter  $\lambda > 0$ ) if

$$\Pr(y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \dots$$

The Poisson Distribution can be derived as a limit of the binomial distribution and will arise in many biological applications involving evolutionary processes.

- (6) Hypergeometric distribution: Jane ran a gene expression experiment where she measured the expression of  $n = 7,000$  genes, of these genes she found that  $n_1 = 500$  were biologically important via a high throughput drug screen. Jane's friend Olaf found  $k = 200$  genes in a toxicity screen. If both the toxicity screen and the drug screen are random and independent with respect to the genes then the probability that there is an overlap of  $k_1$  genes in the drug and toxicity screen lists is

$$\Pr(n_1, n, k_1, k) = \frac{\binom{n_1}{k_1} \binom{n - n_1}{k - k_1}}{\binom{n}{k}}.$$

The classical formulation is with urns. Our urn contains  $n$  balls of which  $n_1$  are black and the remaining  $n - n_1$  are white. We sample  $k$  balls from the urn without replacement. What is the probability that there will be exactly  $k_1$  black balls in the sample  $k$ .

Is the hypergeometric distribution a good idea to model Jane and Olaf's problem ?

### 2.1.3. Means and variances of discrete random variables

**Definition.** *The mean or expectation of a discrete random variable  $Y$  is*

$$\mu = \mathbb{E}[Y] = \sum_y y \Pr(y),$$

where the sum is over all possible values of  $Y$ .

**Example.** Given a binomial random variable  $Y$  its mean is

$$\mu = \mathbb{E}[Y] = \sum_{y=0}^n y \binom{n}{y} p^y (1-p)^{n-y} = np.$$

**Definition.** The mean or expectation of function of discrete random variable  $f(Y)$  is

$$\mathbb{E}[f(Y)] = \sum_y f(y) \Pr(y),$$

where the sum is over all possible values of  $Y$ .

**Definition.** Linearity: If  $Y$  is a discrete random variable and  $\alpha$  and  $\beta$  are constants

$$\mathbb{E}[\alpha + \beta Y] = \sum_y (\alpha + \beta y) \Pr(y) = \alpha + \beta \mu.$$

**Definition.** The variance of a discrete random variable  $Y$  is

$$\sigma^2 = \text{var}[Y] = \sum_y (y - \mu)^2 \Pr(y),$$

where the sum is over all possible values of  $Y$ . The standard deviation is  $\sigma$ .

**Definition.** If  $Y$  is a discrete random variable and  $\alpha$  and  $\beta$  are constants

$$\text{var}[\alpha + \beta Y] = \beta^2 \sigma^2.$$

**Definition.** If  $Y$  is a discrete random variable

$$\sigma^2 = \sum_y y^2 \Pr(y) - \mu^2 = \mathbb{E}(Y^2) - (\mathbb{E}[Y])^2.$$

distribution	mean	variance
Bernoulli	$p$	$p(1-p)$
Binomial	$np$	$np(1-p)$
Uniform	$a + (b-1)/2$	$(b^2-1)/12$
Geometric	$p/(1-p)$	$p/(1-p)^2$
Poisson	$\lambda$	$\lambda$

**Example.** Consider the random variable  $\eta$  that is characterized by the time of the first success of i.i.d. Bernoulli trials  $\xi$

$$\eta = \min\{k \geq 1 : \xi_k = 1\}$$

from our characterization of the geometric distribution

$$\Pr(\eta = k) = (1-p)^{k-1} p, \quad k \geq 1$$

so

$$\mathbb{E}\eta = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = \frac{p}{p^2} = \frac{1}{p}.$$

**Example.** Consider the random variable  $\eta$  that is characterized by the time of first passage of an integer  $N$  of i.i.d. Bernoulli trials  $\xi$ . This means the first time the sum,  $S_n = \sum_{i=1}^n \xi_i$ , equals  $N$

$$\eta = \min\{k \geq 1 : S_k = N\}.$$

The probability distribution for  $\eta$  can be defined recursively

$$\Pr(\eta = k) = p \times \Pr(S_{k-1} = N - 1),$$

and the expectation is

$$\begin{aligned} \mathbb{E}\eta &= p \times \sum_{k=N}^{\infty} k \binom{k-1}{N-1} p^{N-1} (1-p)^{k-N} \\ &= \frac{p^N}{(N-1)!} \sum_{k=N}^{\infty} k(k-1) \cdots (k-N+1) (1-p)^{k-N}. \end{aligned}$$

#### 2.1.4. Moments of distributions and generating functions

**Definition.** If  $Y$  is a discrete random variable then  $\mathbb{E}(Y^r)$  is the  $r$ th moment and

$$\mathbb{E}(Y^r) = \sum_y y^r \Pr(y).$$

**Definition.** If  $Y$  is a discrete random variable then the  $r$ th moment about the mean is moment and

$$\mu_r = \sum_y (y - \mu)^r \Pr(y).$$

The first such moment is zero, the second is the variance  $\sigma^2$  the third is the skewness and the fourth is the kurtosis.

For a discrete random variable the probability generating function pgf will be denoted  $q(t)$  and is defined as

$$q(t) = \mathbb{E}[t^Y] = \sum_y t^y P(y).$$

**Examples.** For the Bernoulli random variable

$$\Pr(y) = p^y (1-p)^{1-y}, \quad y = 0, 1,$$

the pgf is

$$q(t) = 1 - p + pt.$$

For the Bernoulli random variable

$$\Pr(s) = p^{(1+s)/2} (1-p)^{(1-s)/2}, \quad s = -1, 1,$$

the pgf is

$$q(t) = (1-p)t^{-1} + pt.$$

For the binomial random variable

$$\Pr(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 1, 2, \dots, n,$$

the pgf is

$$q(t) = (1-p + pt)^n.$$

The pgf can be used to derive moments of a probability distribution

$$\begin{aligned} \mu &= \left( \frac{d}{dt} q(t) \right)_{t=1} \\ \sigma &= \left( \frac{d^2}{dt^2} q(t) \right)_{t=1} + \mu - \mu^2 \end{aligned}$$

**Theorem.** *If two random variables have pgfs  $q_1(t)$  and  $q_2(t)$  and both pgfs converge in some open interval  $I$  containing 1. If  $q_1(t) = q_2(t)$  for all  $t \in I$ , then the two random variables have the identical distributions.*

### 2.1.5. Continuous random variables

Some random variables are continuous rather than discrete. For example, amount of protein expressed or concentration of an enzyme are continuous valued. Again we use uppercase letters,  $X, Y, Z$ , for the random variable and lowercase letters,  $x, y, z$ , for values the random variable takes. A continuous random variable takes values over a range  $-\infty < X < \infty$  or  $L < X < H$ . Continuous random variables are associated with a probability density function  $p(x)$  for which

$$\Pr(a < X < b) = \int_a^b p(x)dx,$$

and

$$p(x) = \lim_{h \rightarrow 0} \frac{\Pr(x < X < x + h)}{h},$$

and a (cumulative) distribution function

$$F(x) = \int_{-\infty}^x p(u)du,$$

it is clear that  $0 \leq F(x) \leq 1$  and by calculus (the Radon-Nikodym Theorem)

$$p(x) = \frac{d}{dx}F(x).$$

**Definition.** *The mathematical mean or expectation of a random variable  $X$  is*

$$\mu = \mathbb{E}X = \int_{-\infty}^{\infty} xp(x)dx = \int_{-\infty}^{\infty} x dF(x).$$

The mean is in some sense the “center of mass” of the distribution  $F(x)$ . In statistics it is often called a location parameter of the distribution. The following properties are a result of the linearity of expectations and Fubini’s theorem

- (1)  $\mathbb{E}(a + bX) = a + b\mu$
- (2)  $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$
- (3) If  $a \leq X \leq B$  then  $a \leq \mathbb{E}X \leq b$
- (4) The probability of an event  $A$  can be expressed in terms of expectations of an indicator random variable

$$\Pr(A) = \mathbb{E}I(A),$$

where the indicator function  $I(A)$  is 1 if  $A = \text{true}$  and 0 otherwise.

**Definition.** *The variance of a random variable  $X$  is*

$$\sigma^2 = \mathbb{E}(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \min_a \mathbb{E}(X - a)^2.$$

The variance is the amount of dispersion or concentration of the mass of the distribution about the mean. In statistics it is often called a scale parameter of the distribution.

**Definition.** The variance of a random variable  $X$  is

$$\sigma^2 = \mathbb{E}(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \min_a \mathbb{E}(X - a)^2.$$

The second moment of a zero mean random variable is its variance.

**Definition.** The  $r$ -th moment of a random variable  $X$  is

$$\mathbb{E}X^r = \int_{-\infty}^{\infty} x^r p(x) dx.$$

**Definition.** The centered  $r$ -th moment of a random variable  $X$  is

$$\mathbb{E}(X - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r p(x) dx.$$

All the above can be defined for a function  $f(X)$  of a random variable

$$\mathbb{E}f(x) = \mathbb{E}_{x \sim X}[f(x)] = \int_x f(x)p(x)dx.$$

**Example.** Our objective is to diagnose whether a patient has lung cancer type  $A$  or lung cancer type  $B$ . We measure the expression level of 3,000 genes for each patient,  $X \in \mathbb{R}^{3,000}$ . The patients cancer type can be designated as  $Y \in \{A, B\}$ . Patients with lung cancer type  $A$  have density function  $p_A(x) = p(x|A)$  and those with lung cancer type  $B$  have density function  $p_B(x) = P(x|B)$ . The likelihood of developing lung cancer  $A$  and  $B$  are equal. Assume an oracle gives us  $p(x|A)$  and  $p(x|B)$  our goal is to find an optimal classification function and state what its error is. A classification function  $c$  is a map  $c : X \rightarrow \{A, B\}$  and the error of a classification function will be

$$\begin{aligned} \text{Err}[c] &= \mathbb{E}_{x \sim X, y \sim Y}[c(x) \neq y] \\ &= \int_{X, Y} [c(x) \neq y] p(x, y) dx dy \\ &= \int_X ([c(x) \neq A] p(x, A) dx + [c(x) \neq B] p(x, B)) dx, \end{aligned}$$

a classification function that minimizes this error is called a Bayes optimal classifier and this minimum error is called the Bayes error.

Assume an oracle has given us  $p_A(x)$  and  $p_B(x)$  we can now compute the conditional probabilities (likelihoods):

$$\begin{aligned} \text{Pr}(A|x) &= \frac{p(x, A)}{p(x)} = \frac{p(x|A)p(A)}{p(x)}, \\ \text{Pr}(B|x) &= \frac{p(x, B)}{p(x)} = \frac{p(x|B)p(B)}{p(x)}. \end{aligned}$$

We first compute  $p(x)$

$$p(x) = p(x|A)p(A) + p(x|B)p(B),$$

which we can compute since we know  $p(x|A)$  and  $p(x|B)$  and  $p(A) = p(B) = \frac{1}{2}$ .

A the following function  $c^*$  is a Bayes optimal classifier

$$c^*(x) = \begin{cases} A & \text{if } p(A|x) \geq p(B|x), \\ B & \text{o.w.} \end{cases}$$

### 2.1.6. Empirical distributions

In the previous section we defined continuous random variables and the previous example demonstrated how we can characterize important aspects of a problem given a distribution.

However, in most real problems we are not given the distribution  $p(x)$  we are usually given a finite sample of  $n$  observations  $(x_1, \dots, x_n)$  drawn from the distribution  $p(x)$ . We have to make use of the sample values as a proxy for the distribution  $p(x)$ . This is done via the empirical distribution function.

**Definition.** Given a draw  $(x_1, \dots, x_n)$  of  $n$  observations from  $p(x)$  we define the empirical distribution as

$$p_n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x_i),$$

where  $\delta(\cdot)$  is the delta function.

**Definition.** The delta function  $\delta(x)$  is a function with the following property

$$g(x) = \int g(u)\delta(x-u)du,$$

in that it “picks” out a function value at  $x$ .

**Examples.** An example of a delta function is the following

$$\delta(x) = \lim_{\sigma \rightarrow 0} \frac{1}{\sigma\sqrt{2\pi}} e^{-|x|/2\sigma^2}.$$

Another example is

$$\delta(x) = \lim_{a \rightarrow 0} \frac{1}{a} I\left(x \in \left[x - \frac{a}{2}, x + \frac{a}{2}\right]\right).$$

Given the empirical distribution we can define the sample average or average as an expectation.

**Definition.** The sample average or average,  $\hat{\mu}_n$ , of a draw  $(x_1, \dots, x_n)$  is

$$\begin{aligned} \hat{\mu}_n &= \mathbb{E}_n x = \int x p_n(x) dx \\ &= \frac{1}{n} \sum_{i=1}^n \int x \delta(x_i) dx \\ &= \frac{1}{n} \sum_{i=1}^n x_i. \end{aligned}$$

In general we will use quantities like  $\hat{\mu}_n$  as proxy for  $\mu$  and this will be developed in the next section.

**Example.** We revisit example 2.1.5 we were classifying patients as having lung cancer of type A or type B. We formulated the Bayes optimal classifier and the error of the Bayes optimal classifier. In that example we were given  $p(x|A)$  and  $p(x|B)$  by an oracle. In real life we are given in general  $n$  observations  $\{(x_1, A), \dots, (x_n, A)\}$  from  $p(x|A)$  and  $n$  observations  $\{(x_{n+1}, B), \dots, (x_{2n}, B)\}$  from  $p(x|B)$  this defines

an empirical distribution function  $p_{2n}(x, y)$ . We can use this distribution function to measure the empirical error of a classifier  $c$

$$\begin{aligned} \text{Err}_{2n}[c] &= \mathbb{E}_{2n}[c(x) \neq y] \\ &= \int_{X,Y} [c(x) \neq y] p_{2n}(x, y) dx dy \\ &= \frac{1}{2n} \left[ \sum_{i=1}^n [c(x_i) \neq A] + \sum_{i=n+1}^{2n} [c(x_i) \neq B] \right]. \end{aligned}$$

An important question that we will focus on in the next section is how close are  $\text{Err}_{2n}[c]$  and  $\text{Err}[c]$ .

### 2.1.7. Central limit theorems and law of large numbers

The main question addressed in this section is how close is  $\hat{\mu}_n$  to  $\mu$ . We will also motivate why we ask this question.

A sequence of random variables  $y_n$  converges almost surely to a random variable  $Y$  iff  $\mathbb{P}(y_n \rightarrow Y) = 1$ . A sequence of random variables  $y_n$  converges in probability to a random variable  $Y$  iff for every  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} \mathbb{P}(|y_n - Y| > \epsilon) = 0$ . Let  $\hat{\mu}_n := n^{-1} \sum_{i=1}^n x_n$ . The sequence  $x_1, \dots, x_n$  satisfies the strong law of large numbers if for some constant  $c$ ,  $\hat{\mu}_n$  converges to  $c$  almost surely. The sequence  $x_1, \dots, x_n$  satisfies the weak law of large numbers iff for some constant  $c$ ,  $\hat{\mu}_n$  converges to  $c$  in probability. In general the constant  $c$  will be the expectation of the random variable  $\mathbb{E}x$ .

A given function  $f(x)$  of random variables  $x$  concentrates if the deviation between its empirical average,  $n^{-1} \sum_{i=1}^n f(x_i)$  and expectation,  $\mathbb{E}f(x)$ , goes to zero as  $n$  goes to infinity. That is  $f(x)$  satisfies the law of large numbers.

**Theorem (Jensen).** *If  $\phi$  is a convex function then  $\phi(\mathbb{E}x) \leq \mathbb{E}\phi(x)$ .*

**Theorem (Bienaymé-Chebyshev).** *For any random variable  $x$ ,  $\epsilon > 0$*

$$\Pr(|x| \geq \epsilon) \leq \frac{\mathbb{E}x^2}{\epsilon^2}.$$

*Proof.*

$$\mathbb{E}x^2 \geq \mathbb{E}(x^2 I(|x| \geq \epsilon)) \geq \epsilon^2 \Pr(|x| \geq \epsilon). \quad \square$$

**Theorem (Chebyshev).** *For any random variable  $x$ ,  $\epsilon > 0$*

$$\Pr(|x - \mu| \geq \epsilon) \leq \frac{\text{Var}(x)}{\epsilon^2}.$$

*Proof.* Same as above.  $\square$

**Theorem (Markov).** *For any random variable  $x$ ,  $\epsilon > 0$*

$$\Pr(|x| \geq \epsilon) \leq \frac{\mathbb{E}e^{\lambda x}}{e^{\lambda \epsilon}}$$

and

$$\Pr(|x| \geq \epsilon) \leq \inf_{\lambda < 0} e^{-\lambda \epsilon} \mathbb{E}e^{\lambda x}.$$

*Proof.*

$$\Pr(x > \epsilon) = \Pr(e^{\lambda x} > e^{\lambda \epsilon}) \leq \frac{\mathbb{E}e^{\lambda x}}{e^{\lambda \epsilon}}. \quad \square$$

For the sums or averages of independent random variables the above bounds can be improved from polynomial in  $1/\epsilon$  to exponential in  $\epsilon$ .

The following theorems will be for zero mean random variables. The extension to nonzero mean is trivial.

**Theorem** (Bennet). *Let  $x_1, \dots, x_n$  be independent random variables with  $\mathbb{E}x = 0$ ,  $\mathbb{E}x^2 = \sigma^2$ , and  $|x_i| \leq M$ . For  $\epsilon > 0$*

$$\Pr\left(\left|\sum_{i=1}^n x_i\right| > \epsilon\right) \leq 2e^{-\frac{n\sigma^2}{M^2}\phi\left(\frac{\epsilon M}{n\sigma^2}\right)},$$

where

$$\phi(z) = (1+z)\log(1+z) - z.$$

*Proof.* We will prove a bound on one-side of the above theorem

$$\Pr\left(\sum_{i=1}^n x_i > \epsilon\right).$$

$$\begin{aligned} \Pr\left(\sum_{i=1}^n x_i > \epsilon\right) &\leq e^{-\lambda \epsilon} \mathbb{E}e^{\lambda \sum x_i} = e^{-\lambda \epsilon} \prod_{i=1}^n \mathbb{E}e^{\lambda x_i} \\ &= e^{-\lambda \epsilon} (\mathbb{E}e^{\lambda x})^n. \end{aligned}$$

$$\begin{aligned} \mathbb{E}e^{\lambda x} &= \mathbb{E}\sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} = \sum_{k=0}^{\infty} \lambda^k \frac{\mathbb{E}x^k}{k!} \\ &= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}x^2 x^{k-2} \leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} M^{k-2} \sigma^2 \\ &= 1 + \frac{\sigma^2}{M^2} \sum_{k=2}^{\infty} \frac{\lambda^k M^k}{k!} = 1 + \frac{\sigma^2}{M^2} (e^{\lambda M} - 1 - \lambda M) \\ &\leq e^{\frac{\sigma^2}{M^2} (e^{\lambda M} - \lambda M - 1)}. \end{aligned}$$

The last line holds since  $1 + x \leq e^x$ .

Therefore,

$$(2.1) \quad \Pr\left(\sum_{i=1}^n x_i > \epsilon\right) \leq e^{-\lambda \epsilon} e^{\frac{\sigma^2}{M^2} (e^{\lambda M} - \lambda M - 1)}.$$

We now optimize with respect to  $\lambda$  by taking the derivative with respect to  $\lambda$

$$\begin{aligned} 0 &= -\epsilon + \frac{n\sigma^2}{M^2} (Me^{\lambda M} - M), \\ e^{\lambda M} &= \frac{\epsilon M}{n\sigma^2} + 1, \\ \lambda &= \frac{1}{M} \log\left(1 + \frac{\epsilon M}{n\sigma^2}\right). \end{aligned}$$

The theorem is proven by substituting  $\lambda$  into equation (2.1).  $\square$

The problem with Bennet's inequality is that it is hard to get a simple expression for  $\epsilon$  as a function of the probability of the sum exceeding  $\epsilon$ .

**Theorem** (Bernstein). *Let  $x_1, \dots, x_n$  be independent random variables with  $\mathbb{E}x = 0$ ,  $\mathbb{E}x^2 = \sigma^2$ , and  $|x_i| \leq M$ . For  $\epsilon > 0$*

$$\Pr \left( \left| \sum_{i=1}^n x_i \right| > \epsilon \right) \leq 2e^{-\frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}}.$$

*Proof.*

Take the proof of Bennet's inequality and notice

$$\phi(z) \geq \frac{z^2}{2 + \frac{2}{3}z}. \quad \square$$

**Remark.** With Bernstein's inequality a simple expression for  $\epsilon$  as a function of the probability of the sum exceeding  $\epsilon$  can be computed

$$\sum_{i=1}^n x_i \leq \frac{2}{3}uM + \sqrt{2n\sigma^2u}.$$

*Outline.*

$$\Pr \left( \sum_{i=1}^n x_i > \epsilon \right) \leq 2e^{-\frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}} = e^{-u},$$

where

$$u = \frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}.$$

we now solve for  $\epsilon$

$$\epsilon^2 - \frac{2}{3}\epsilon M - 2n\sigma^2\epsilon = 0$$

and

$$\epsilon = \frac{1}{3}uM + \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2u}.$$

Since  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$

$$\epsilon = \frac{2}{3}uM + \sqrt{2n\sigma^2u}.$$

So with large probability

$$\sum_{i=1}^n x_i \leq \frac{2}{3}uM + \sqrt{2n\sigma^2u}. \quad \triangle$$

If we want to bound

$$|\hat{\mu}_n - \mu| = |n^{-1} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x)|$$

we consider

$$|f(x_i) - \mathbb{E}f(x)| \leq 2M.$$

Therefore

$$\sum_{i=1}^n (f(x_i) - \mathbb{E}f(x)) \leq \frac{4}{3}uM + \sqrt{2n\sigma^2u}$$

and

$$n^{-1} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x) \leq \frac{4uM}{3n} + \sqrt{\frac{2\sigma^2 u}{n}}.$$

Similarly,

$$\mathbb{E}f(x) - n^{-1} \sum_{i=1}^n f(x_i) \geq \frac{4uM}{3n} + \sqrt{\frac{2\sigma^2 u}{n}}.$$

In the above bound

$$\sqrt{\frac{2\sigma^2 u}{n}} \geq \frac{4uM}{n}$$

which implies  $u \leq \frac{n\sigma^2}{8M^2}$  and therefore

$$|n^{-1} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x)| \lesssim \sqrt{\frac{2\sigma^2 u}{n}} \text{ for } u \lesssim n\sigma^2,$$

which corresponds to the tail probability for a Gaussian random variable and is predicted by the Central Limit Theorem (CLT) Condition that  $\lim_{n \rightarrow \infty} n\sigma^2 \rightarrow \infty$ . If  $\lim_{n \rightarrow \infty} n\sigma^2 = C$ , where  $C$  is a fixed constant, then

$$|n^{-1} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x)| \lesssim \frac{C}{n}$$

which corresponds to the tail probability for a Poisson random variable.

**Proposition.** *Bounds between the difference of sums of two draws of a random variable can be derived in the same manner as for the deviation between the sum and difference. This is called symmetrization in empirical process theory.*

*Given two draws from a distribution  $x_1, \dots, x_n$  and  $x'_1, \dots, x'_n$*

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \frac{1}{n} \sum_{i=1}^n f(x'_i) \right| \geq \epsilon \right) \leq e^{-\frac{\epsilon^2 n}{8(2\sigma^2 + 2M\epsilon/2)}}.$$

*Proof.* From Bernstein's inequality we know that

$$\begin{aligned} \Pr \left( \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x) \right| \geq \epsilon \right) &\leq 2e^{-\frac{\epsilon^2 n}{2\sigma^2 + 2M\epsilon/2}}, \\ \Pr \left( \left| \frac{1}{n} \sum_{i=1}^n f(x'_i) - \mathbb{E}f(x) \right| \geq \epsilon \right) &\leq 2e^{-\frac{\epsilon^2 n}{2\sigma^2 + 2M\epsilon/2}}. \end{aligned}$$

If we can ensure

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x) \right| &\leq \epsilon/2 \\ \left| \frac{1}{n} \sum_{i=1}^n f(x'_i) - \mathbb{E}f(x) \right| &\leq \epsilon/2, \end{aligned}$$

then

$$\left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \frac{1}{n} \sum_{i=1}^n f(x'_i) \right| \leq \epsilon.$$

Applying Bernstein's inequality to the first two conditions gives us the desired result.  $\square$

**Proposition.** *In addition to bounds on sums of random variables for certain distributions there also exist approximations to these sums. Let  $x_1, \dots, x_n$  be independent random variables with  $\mathbb{E}x = 0$ ,  $\mathbb{E}x^2 = \sigma^2$ , and  $|x_i| \leq M$ . For  $\epsilon > 0$*

$$\Pr \left( \left| \sum_{i=1}^n x_i \right| > \epsilon \right) \approx \Phi(\epsilon, \sigma, n, M).$$

*The approximations are computed using similar ideas as those in computing the upper bounds.*

**Example.** *Hypothesis testing is an example where limit distributions are applied. Assume we are given  $n$  observations from two experimental conditions  $x_1, \dots, x_n$  and  $z_1, \dots, z_n$ .*

*We ask the following question: are the means of the observations of the two experimental conditions the same ?*

*This is formulated as: can we reject the null hypothesis that the difference of the means of the two two experimental conditions the same ?*

*If  $x_1, \dots, x_n$  and  $z_1, \dots, z_n$  were drawn according to a distribution  $p(u)$  we would know that*

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n z_i \right| \geq \epsilon \right) \leq e^{-\frac{\epsilon^2 n}{8(2\sigma^2 + 2M\epsilon/2)}}.$$

*So given the difference*

$$d = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n z_i.$$

*We can bound the probability this would occur by chance under the null hypothesis. Thus if  $d$  is very large the chance that the null hypothesis gave rise to  $d$  is small and so we may reject the null hypothesis.*

### 2.1.8. My favorite continuous distributions

- (1) Uniform distribution: Given an interval  $I = [a, b]$  then

$$p(x) = \frac{1}{a-b} \quad \text{for } x \in I.$$

The mean and variance are  $\mu = \frac{a+b}{2}$  and  $\sigma^2 = \frac{(a-b)^2}{12}$ . If the interval  $I = [0, 1]$  then  $F(x) = x$ . Any continuous distribution can be transformed into the uniform distribution via a monotonic function. A monotonic function preserves ranks. For this reason, study of rank statistics reduces to the study of properties of the uniform distribution.

- (2) Gaussian or normal distribution: The random variable  $x \in (-\infty, \infty)$  with

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where  $\mu$  is the mean and  $\sigma^2$  is the variance. It is often written as  $N(\mu, \sigma)$ . The case of  $N(0, 1)$  is called the standard normal in that the normal distribution has been standardized. Given a draw from a normal distribution  $x \sim N(\mu, \sigma)$  the z-score transforms this to a standard normal random variable

$$z = \frac{x - \mu}{\sigma}.$$

We stated before that the probability distribution of a binomial random variable with parameters  $p$  and  $n$  trials is

$$\Pr(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 1, 2, \dots, n.$$

There is no closed form formulation of the above distribution. However, for large enough  $n$  (say  $n > 20$ ) and  $p$  not too near 0 or 1 (say  $0.05 < p < 0.95$ ) we can approximate the binomial as a Gaussian. To fix the x-axis we normalize the above from number of successes  $y$  to the ratio of success  $\frac{y}{n}$

$$\Pr\left(\frac{y}{n}\right) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 1, 2, \dots, n.$$

The above random variable can be approximated by  $N\left(p, \frac{p(1-p)}{n}\right)$ . This is the normal approximation of the Binomial distribution.

We will often use continuous distributions to approximate discrete ones.

- (3) Exponential distribution: The random variable  $X$  takes the possible values  $X \in [0, \infty)$

$$p(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0,$$

the distribution function is

$$F(x) = 1 - e^{-\lambda x},$$

and  $\mu = 1/\lambda$  and  $\sigma^2 = 1/\lambda^2$ .

The exponential distribution can be used to approximate the geometric distribution

$$\begin{aligned} p(y) &= (1-p)p^y \quad \text{for } y = 1, 2, \dots \\ F(y) &= 1 - p^{-y+1}. \end{aligned}$$

If we set the random variable  $Y = \lfloor X \rfloor$  then

$$\begin{aligned} \Pr(Y = y) &= \Pr(y \leq x \leq y+1) \\ &= (1 - e^{-\lambda})e^{-\lambda y}, \quad \text{for } y = 1, 2, \dots \end{aligned}$$

so we have  $p = e^{-\lambda}$  and  $\mu = 1/(e^{-\lambda} - 1)$  and  $\sigma^2 = e^\lambda/(e^\lambda - 1)^2$ .

- (4) Gamma Distribution: The exponential distribution is a special case of the gamma distribution

$$p(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} \quad x > 0,$$

where  $\lambda, k > 0$  and  $\Gamma(k)$  is the gamma function

$$\Gamma(u) = \int_0^\infty e^{-t} t^{u-1} dt.$$

When  $u$  is an integer

$$\Gamma(u) = (u-1)!.$$

The mean and variance are  $\mu = k/\lambda$  and  $\sigma^2 = k/\lambda^2$ . The exponential distribution is a gamma distribution with  $k = 1$ . The chi-square distribution,

with  $\nu$  degrees of freedom, is another example of a gamma distribution with  $\lambda = 1/2$  and  $k = \nu/2$

$$p(x) = \frac{\lambda^k x^{\nu/2-1} e^{-x/2}}{2^{\nu/2} \Gamma(\nu/2)}$$

when  $\nu = 1$  the chi-square distribution is the distribution of  $x^2$  where  $X \sim N(0, 1)$ .

- (5) Beta Distribution: A continuous random variable  $X$  has a beta distribution if for  $\alpha, \beta > 0$

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad 0 < x < 1.$$

The mean and variance are  $\mu = \alpha/(\alpha + \beta)$  and  $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ .

The uniform distribution is a special case of the beta distribution with  $\alpha, \beta = 1$ . The beta distribution will arise often in Bayesian statistics as a prior that is somewhat spread out.

### 2.1.9. Moment generating functions

Moment generating functions become useful in analyzing sums of random variables as well random walks and extremal statistics.

**Definition.** *The moment generating function of a continuous random variable  $X$  is*

$$M(t) = \mathbb{E}[e^{tx}] = \int_L^H e^{tx} p(x) dx = \int_L^H e^{tx} dF(x),$$

if for some  $\delta > 0$ ,  $M(t) < \infty$  for  $t \in (-\delta, \delta)$ .

The domain of support of  $M$  is

$$D = \{t : M(t) < \infty\}.$$

We will require that  $\delta > 0$  since we will require  $M$  to have derivatives

$$\begin{aligned} M'(t) &= \frac{dM(t)}{dt} \\ &= \frac{d\mathbb{E}(e^{tx})}{dt} \\ &= \mathbb{E} \frac{d(e^{tx})}{dt} \\ &= \mathbb{E} [xe^{tx}]. \end{aligned}$$

So  $M'(0) = \mathbb{E}[x]$  similarly  $M^k(t) = \mathbb{E}[x^k e^{tx}]$  and the moment generating function can be used to generate moments.

$$M^{(k)}(0) = \mathbb{E}x^k.$$

Another consequence of this is that for  $\delta > 0$  we have a power (McLaurin) series expansion

$$\begin{aligned} M(t) &= \mathbb{E}[e^{tx}] \\ &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(tx)^k}{k!}\right] \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}x^k \end{aligned}$$

The derivatives of the coefficients in the power series of  $M$  about zero give us the  $k$ -th moments about zero.

### 2.1.10. Statistics and inference

Statistics is in some sense inverse probability. In probability we characterized properties of particular trials and distributions and computed summary statistics such as means, moments, and variances.

In statistics we consider the inverse problem:

given an i.i.d. draw of  $n$  samples  $(x_1, \dots, x_n)$  from a distribution  $p(x)$

- (1) What is  $p(x)$  or what are summary statistics of  $p(x)$  ?
- (2) Are the summary statistics sufficient to characterize  $p(x)$  ?
- (3) How accurate is our estimate of  $p(x)$  ?

There are a multitude of ways of addressing the above questions (much like “The Library of Babel”). This keeps statisticians employed.

We first focus on two methods (that are often result in similar results):

- Maximum likelihood (ML)
- Statistical decision theory (SDT).

We will then see why for some inference problems we need to constrain statistical models, often by using priors (Bayesian methods).

In general statistical models come in two flavors: parametric and nonparametric. We will focus on parametric models in this introduction but in the classification section of the course will also look at nonparametric models.

We first look at the ML framework. We start with a model

$$p(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

generally Statisticians designate parameters in models as  $\theta$  and the domain as  $\Theta$ , in this case  $\theta = \{\mu, \sigma\}$ . Given an i.i.d. draw of  $n$  samples from  $p(x|\theta)$ ,  $D = (x_1, \dots, x_n) \sim p(x|\theta)$  it is natural to examine the likelihood function  $p(D|\theta)$  and find the  $\theta \in \Theta$  that maximizes this. This is maximum likelihood estimation (MLE)

$$\max_{\theta \in \Theta} \left[ p(D|\theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2} \right].$$

the parameters that maximize the likelihood will also maximize the log-likelihood, since log is a monotonic function in its argument

$$L(D|\theta) = \log p(D|\theta)$$

and

$$\hat{\theta} = \arg \max_{\theta} \left[ L = \sum_{i=1}^n -\frac{\log(\sigma^2 2\pi)}{2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right],$$

where  $\hat{\theta}$  is our estimate of the model parameter. We can solve for both  $\sigma, \mu$  using calculus

$$\begin{aligned} \frac{dL}{d\mu} &= 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} \\ \frac{dL}{d\sigma} &= 0 \Rightarrow \hat{\sigma} = n^{-1} \sum_{i=1}^n (x_i - \hat{\mu})^2. \end{aligned}$$

In explaining the SDT framework and looking more carefully at the ML framework we will use (linear) regression as our example. The problem of linear regression is formulated as follows we are given  $n$  pairs of samples  $D = \{(x_1, y, 1), \dots, (x_n, y_n)\}$  with independent variables  $x \in \mathbb{R}^n$  and dependent variables  $y \in \mathbb{R}$  and we would like to estimate the functional relation of  $x$  to  $y$   $f : x \rightarrow y$ .

We start with linear regression in one dimension, a problem addressed by Gauss and Legendre. So  $x, y \in \mathbb{R}$  and our model or class of functions are linear functions

$$f_{\theta} = \theta x.$$

In the SDT framework we need to select not only a model but a loss function. We will use square loss as our loss function

$$L(f(x), y) = (f(x) - y)^2.$$

If we are given the joint distribution  $p(x, y)$ , then we can solve

$$\min_{\theta} \mathbb{E}(Y - \theta X)^2,$$

minimizing the above results in

$$\theta = \frac{\mathbb{E}XY}{\mathbb{E}XX}$$

and if  $\mathbb{E}x = 0$

$$\theta = \frac{\mathbb{E}XY}{\sigma^2}.$$

However, we are not give the joint  $p(x, y)$  or the marginal  $p(x)$  or the conditional  $p(y|x)$  distributions. What we are given is a draw of  $n$  observations from the joint  $D = \{(x_1, y_1), \dots, (x_n, y_n)\} \sim p(x, y)$ . We replace the expected error

$$\mathbb{E}(Y - \theta X)^2,$$

with the empirical error

$$\mathbb{E}_n(Y - \theta X)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \theta x_i)^2,$$

and hope that minimizing the empirical error will result in a solution that is close to the minima of the expected error. Minimizing

$$\text{Err}_n = \frac{1}{n} \sum_{i=1}^n (y_i - \theta x_i)^2,$$

can be done by taking the derivative and setting it to zero

$$\begin{aligned}\frac{d\text{Err}_n}{d\theta} &= 0 = -\frac{2}{n} \sum_{i=1}^n (y_i - \theta x_i) x_i, \\ &= -\frac{2}{n} \sum_{i=1}^n y_i x_i + \frac{2}{n} \sum_{i=1}^n \theta x_i^2,\end{aligned}$$

which implies

$$\hat{\theta} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}.$$

We now solve the same problem from the ML framework. Again  $n$  observations from the joint distribution  $D = \{(x_1, y_1), \dots, (x_n, y_n)\} \sim p(x, y)$ . We assume that

$$y = \theta x + \varepsilon,$$

where  $\varepsilon \sim N(0, \sigma^2)$  is Gaussian. This implies that

$$y - \theta x \sim N(0, \sigma^2)$$

and our likelihood can be modeled as

$$p(D|\theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - \theta x_i)^2 / 2\sigma^2}.$$

We now maximize the log-likelihood  $L(D|\theta)$

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \left[ -\frac{\log(\sigma^2 2\pi)}{2} - \frac{(y_i - \theta x_i)^2}{2\sigma^2} \right].$$

Taking the derivative and setting it to zero

$$\frac{L(D|\theta)}{d\theta} = 0 = \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta x_i) x_i,$$

which implies as before

$$\hat{\theta} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}.$$

This implies that there is a relation between using a square loss in the SDT framework and Gaussian model of noise in the ML framework, This relation can be formalized with more careful analysis.

Both ML and SDT as formulated work very well for the above problems and by this I mean for  $n > 20$  with very high probability we estimate the right slope in the model  $\hat{\theta} \approx \theta$ .

We now examine a case where both methods fail. This case also happens to be a common a problem in many genomic applications. We again are looking at the problem of linear regression with the dependent variable  $y \in \mathbb{R}$  is for example blood pressure or the concentration of a specific such as Prostate Specific Antigen (PSA), the independent variables  $x \in \mathbb{R}^p$  are measurements over the genome, for example expression levels for genes, so  $p$  is large (typically 7,000 – 50,000). We draw  $n$  samples  $D = \{(x_1, y_1), \dots, (x_n, y_n)\} \sim p(x, y)$  from the joint. Typically,  $n$  is not so large, 30 – 200. We would like to estimate  $\hat{\theta} \in \mathbb{R}^p$ . In statistics this is called the large  $p$  small  $n$  problem and we will see it causes problems for the standard methods we have discussed. In the computer science/machine learning literature

the problem is called learning high-dimensional data and the number of samples is associated with the variable  $m$  and the dimensions with  $n$ .

We first look at the SDT formulation. We would like minimize with respect to

$$L = \sum_{i=1}^n (y_i - x_i \cdot \theta)^2.$$

We will rewrite the above in matrix notation. In doing this we define a matrix  $\mathbf{X}$  which is  $n \times p$  and each row of the matrix is a data point  $x_i$ . We also define a column vector  $y$  ( $p \times 1$ ) with  $y_i$  as the  $i$ -th element of  $y$ . Similarly  $\theta$  is a column vector with  $p$  rows. We can rewrite the error minimization as

$$\arg \min_{\theta} [L = (y - \mathbf{X}\theta)^T (y - \mathbf{X}\theta)],$$

taking derivatives with respect to  $\theta$  and setting this equal to zero (taking derivatives with respect to  $\theta$  means taking derivatives with respect to each element in  $\theta$ )

$$\begin{aligned} \frac{dL}{d\theta} &= -2\mathbf{X}^T (y - \mathbf{X}\theta) = 0 \\ &= \mathbf{X}^T (y - \mathbf{X}\theta) = 0. \end{aligned}$$

this implies

$$\begin{aligned} \mathbf{X}^T y &= \mathbf{X}^T \mathbf{X} \theta \\ \hat{\theta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y. \end{aligned}$$

If we look at the above formula carefully there is a serious numerical problem –  $(\mathbf{X}^T \mathbf{X})^{-1}$ . The matrix  $\mathbf{X}^T \mathbf{X}$  is a  $p \times p$  matrix of rank  $n$  where  $p \gg n$ . This means that  $\mathbf{X}^T \mathbf{X}$  cannot be inverted so we cannot compute the estimate  $\hat{\theta}$  by matrix inversion. There are numerical approaches to address this issue but the solution will not be unique or stable. A general rule in estimation problems is that numerical problems in estimating the parameters usually coincide with with statistical errors or variance of the estimate.

We now solve the above case in the ML framework

$$y = \theta \cdot x + \varepsilon,$$

where  $\varepsilon \sim N(0, \sigma^2)$  is Gaussian. This implies that

$$y - \theta \cdot x \sim N(0, \sigma^2)$$

and our likelihood can be modeled as

$$\begin{aligned} p(D|\theta) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - \theta \cdot x_i)^2 / 2\sigma^2} \\ &= \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{\sum_{i=1}^n -(y_i - \theta \cdot x_i)^2 / 2\sigma^2} \\ &= \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-(y - \mathbf{X}\theta)^T (y - \mathbf{X}\theta) / 2\sigma^2}. \end{aligned}$$

If we maximize the log-likelihood we again have the formula

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y.$$

In the SDT framework the large  $p$  small  $n$  problem is often dealt with by using a penalized loss function instead to the empirical loss function. This is often called a shrinkage estimator or regularization. The following is a commonly used estimator

$$L = \sum_{i=1}^n (y_i - x_i \cdot \theta)^2 + \lambda \|\theta\|^2,$$

where  $\lambda > 0$  is a parameter of the problem and  $\|\theta\|^2 = \sum_{i=1}^p \theta_i^2$ .

$$\begin{aligned} \frac{dL}{d\theta} &= -2\mathbf{X}^T(y - \mathbf{X}\theta) + 2\lambda\theta = 0 \\ &= \mathbf{X}^T(\mathbf{X}\theta - y) + \lambda\theta. \end{aligned}$$

this implies

$$\begin{aligned} \mathbf{X}^T y &= \mathbf{X}^T \mathbf{X} \theta + \lambda \theta \\ \mathbf{X}^T y &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \theta \\ \hat{\theta} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T y, \end{aligned}$$

where  $\mathbf{I}$  is the  $p \times p$  identity matrix. The matrix  $(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})$  is invertible and this penalized loss function approach has had great success in large  $p$  small  $n$  problems.

We will now reframe the penalized loss approach in a probabilistic framework which will result in going from the ML approach to a Bayesian approach. Before we do this we first state Bayes rule

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)}, \\ p(\theta|D) &\propto p(D|\theta)p(\theta), \end{aligned}$$

where  $p(D|\theta)$  is the likelihood we used in the ML method,  $p(\theta)$  is a prior on the models  $\theta$  we place before seeing the data,  $p(D)$  is the probability of the data which we consider a constant, and  $p(\theta|D)$  is the posterior probability of  $\theta$  after observing the data  $D$ . This posterior probability is the key quantity used in Bayesian inference.

We will now rewrite the penalized loss approach such that it can be interpreted as a posterior. We can state our prior probability on  $\theta$  as

$$p(\theta) \propto e^{-\lambda \|\theta\|^2}.$$

Our likelihood can be given by a Gaussian noise model on the error

$$\begin{aligned} p(D|\theta) &\propto \prod_{i=1}^n e^{-(y_i - \theta \cdot x_i)^2 / 2\sigma^2} \\ &\propto e^{-\sum_{i=1}^n (y_i - \theta \cdot x_i)^2 / 2\sigma^2}. \end{aligned}$$

The posterior is

$$\begin{aligned} p(\theta|D) &\propto e^{-\sum_{i=1}^n (y_i - \theta \cdot x_i)^2 / 2\sigma^2} \times e^{-\lambda \|\theta\|^2}, \\ \text{posterior} &\propto \text{likelihood} \times \text{prior}. \end{aligned}$$

We now have two choices before us:

- (1) maximize the posterior with respect to  $\theta$ , results in  $\hat{\theta}$  being a maximum a posteriori (MAP) estimate
- (2) simulate the full posterior and use statistics of this to give us an estimator.

## LECTURE 3

### Hypothesis testing

The framework of hypothesis testing is used in computational biology extensively.

We will look at hypothesis testing in both the classical and Bayesian framework as well as multiple hypothesis testing.

The Bayesian and classical frameworks ask two different questions:

- Bayesian: “What is the probability of the hypothesis being true given data ?” -  $\Pr(H = t|D)$ , posterior.
- Classical: “Assume the hypothesis is true, what is the probability of the data?” -  $\Pr(D|H = t)$ , likelihood.

The first question is more natural but requires a prior on hypotheses.

#### 3.0.11. Classical hypothesis testing

The classical hypothesis testing formulation is called the Neyman-Pearson paradigm. It is a formal means of distinguishing between probability distributions based upon random variables generated from one of the distributions.

The two distributions are designated as:

- the null hypothesis:  $H_0$
- the alternative hypothesis:  $H_A$ .

There are two types of hypothesis, simple and composite:

- simple hypothesis: all aspects of the distribution are specified. For example,  $H_0 : X \sim N(\mu_1, \sigma^2)$  is simple since the distribution is fully specified. Similarly  $H_A : X \sim N(\mu_2, \sigma^2)$  is also simple.
- composite hypotheses: the hypothesis does not specify the distribution. For example,  $H_A : X \sim \text{Bernoulli}(n, p > .25)$ , is composite since  $p > .25$  in the Bernoulli distribution so the distribution is not specified.

In general two types of the alternative hypothesis are one and two sided:

- one sided:  $H_A : X \sim \text{Binomial}(n, p > /25)$
- two sided:  $H_A : X \sim \text{Binomial}(n, p \neq /25)$ .

In this paradigm we first need a test statistic  $t(\mathbf{X})$  which can be computed from the data  $\mathbf{X} = (x_1, \dots, x_n)$ . We have a decision problem in that given  $\mathbf{X}$  we compute  $t(\mathbf{X})$  and decide whether we reject  $H_0$ , a positive event, or accept  $H_0$  the negative

event. The sets of values of  $t$  for which we accept  $H_0$  is the acceptance region and the sets for which we reject  $H_0$  are the rejection region. In this paradigm the following four events written in a contingency table can happen. Two of the events are errors,  $B$  and  $C$ .

	$H_0 = T$	$H_A = T$
Accept $H_0$	A	B
Reject $H_0$	C	D

$C$  is called a type I error and measure the probability of a false positive,

$$\alpha = \Pr(\text{null is rejected when true}).$$

The reason why it is called a false positive is that rejecting the null is a positive since one in general in an experiment is looking to reject the null since this corresponds to finding something different from the lack of an effect. In general in the hypothesis testing framework we will control the  $\alpha$  value explicitly (this will be our knob) and is called the significance level.

$B$  is called the type II error and measure the probability of false negatives,

$$\beta = \Pr(\text{null is accepted when it is false}).$$

The power of a test is

$$1 - \beta = \Pr(\text{null is rejected when it is false}).$$

Ideally we would like a test with  $\alpha = \beta = 0$ . Like most of life this ideal is impossible. For a fixed sample size increasing  $\alpha$  will in general decrease  $\beta$  and vice versa. So the general prescription in designing hypothesis tests is to fix  $\alpha$  of a small number and design a test that will give as small a  $\beta$  as possible, the most powerful test.

**Example.** We return to the DNA sequence matching problem where we get a string of  $2n$  letters corresponding to two strands of DNA ask about the significance of the number of observed matches. Our null hypothesis is that the nucleotides  $A, C, T, G$  are equally likely and independent. Another way of saying this is

$$H_0 : X \sim \text{Binomial}(p = .25, n)$$

the alternative hypothesis is

$$H_A : X \sim \text{Binomial}(p > .25, n).$$

Assume we observe  $Y = 32, 33$  matches out of 100 according to the distribution under the null hypothesis.

$$\Pr(Y \geq 32 | p = .25, n = 100) = .069,$$

$$\Pr(Y \geq 33 | p = .25, n = 100) = .044.$$

Therefore, to achieve an  $\alpha$  level of .05 we would need a significance point (critical value) of 33.

The p-value is the smallest  $\alpha$  for which the null will be rejected. It is also called the achieved significance level. Another way of stating this is the p-value is the probability of obtaining an observed value under the distribution given by the

null hypothesis that is greater (more extreme) than the statistic computed on the data  $t(\mathbf{X})$ .

**Example.** We return to the matching problem.

$$H_0 : X \sim \text{Binomial}(p = .25, n)$$

the alternative hypothesis is

$$H_A : X \sim \text{Binomial}(p > .25, n).$$

We find 11 matches out of 26

$$\Pr(Y \geq 11 | p = .25, n = 26) = .04,$$

so the  $p$ -value is .04

We find 278 matches out of 100

$$\Pr(Y \geq 278 | p = .25, n = 100) \approx .022,$$

and by the Normal approximation of the Binomial

$$Y \sim N(250, 187.5).$$

We now look at an example that introduces a classic null distribution, the  $t$ -statistic and the  $t$ -distribution.

**Example.** We have two cell types cancer A and B. we measure the expression of one protein from the two cells.

$$A : X_1, \dots, X_m \sim N(\mu_1, \sigma^2)$$

$$B : Y_1, \dots, Y_n \sim N(\mu_2, \sigma^2)$$

so we draw  $m$  observations from cell type A and  $n$  from cell type B.

$$H_0 : \mu_1 = \mu_2 = \mu$$

$$H_A : \mu_1 \neq \mu_2.$$

The statistic we use is the  $t$ -statistic

$$t(\mathbf{X}) = \frac{(\bar{X} - \bar{Y})\sqrt{mn}}{\hat{S}\sqrt{m+n}},$$

where

$$\hat{S}^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n-2}.$$

The distribution under the null hypothesis of the above is the  $t$ -distribution with  $d = m + n - 2$  degrees of freedom

$$p(t) = \frac{\Gamma[(d+1)/2]}{\sqrt{d\pi}\Gamma[d/2](1 + \frac{t^2}{d})^{(d+1)/2}}.$$

The above example is the most classical example of a hypothesis test and statistic. It is also a parametric test in that we have a parametric assumption for the null hypothesis. Here the assumption is that the samples are normally distributed.

There are a class of hypothesis tests that are called nonparametric in that the parametric assumptions on the null hypothesis are weak. Typically these tests are

called rank statistics in that the ranks of the observations are used to compute the statistic. We will look at two such statistics: the Mann-Whitney (MW) and Kolmogorov-Smirnov statistics. We first define the MW statistic and state its property. We then look more carefully at the KS statistic and use it to illustrate why rank based statistics are nonparametric.

**Example** (Mann-Whitney statistic). *We have two cell types cancer A and B. we measure the expression of one protein from the two cells.*

$$\begin{aligned} A & : X_1, \dots, X_m \sim F_A \\ B & : Y_1, \dots, Y_n \sim F_B, \end{aligned}$$

where  $F_A$  and  $F_B$  are continuous distributions. This is why the test is nonparametric.

$$\begin{aligned} H_0 & : \mu_1 = \mu_2 = \mu \\ H_A & : \mu_1 > \mu_2. \end{aligned}$$

We first combine the lists

$$Z = \{X_1, \dots, X_m, Y_1, \dots, Y_n\},$$

we then rank order  $Z$

$$Z_{(r)} = \{Z_{(1)}, \dots, Z_{(m+n)}\}.$$

Given the rank ordered list  $Z_{(r)}$  we can compute two statistics

$$\begin{aligned} R_1 & = \text{sum of ranks of samples in A in } Z_{(r)} \\ R_2 & = \text{sum of ranks of samples in B in } Z_{(r)} \end{aligned}$$

Given  $R_1$  and  $R_2$  we compute the following statistics

$$\begin{aligned} U_1 & = mn + \frac{(m+1)m}{2} - R_1 \\ U_2 & = mn + \frac{(n+1)n}{2} - R_2, \end{aligned}$$

$U = \min(U_1, U_2)$ . The statistic

$$\hat{z} = \frac{|U - \frac{mn}{2}|}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0, 1).$$

This test is called nonparametric in the the distributional assumptions (in terms of parameters) on the null hypothesis are extremely weak.

**Example** (Kolmogorov-Smirnov statistic). *We have two cell types cancer A and B. we measure the expression of one protein from the two cells.*

$$\begin{aligned} A & : X_1, \dots, X_m \sim F_A \\ B & : Y_1, \dots, Y_n \sim F_B, \end{aligned}$$

where  $F_A$  and  $F_B$  are continuous distributions. This is why the test is nonparametric.

$$\begin{aligned} H_0 & : F_A = F_B \\ H_A & : F_A \neq F_B. \end{aligned}$$

We first construct empirical distribution functions for the two sets of data  $X = \{X_1, \dots, X_m\}$ ,  $Y = \{Y_1, \dots, Y_n\}$

$$F_m(x) = \frac{\#\{X \leq x\}}{m}$$

$$F_n(x) = \frac{\#\{Y \leq x\}}{n},$$

where  $\#\{X \leq x\}$  indicates the number of elements in  $X$  that are smaller than  $x$ , similarly for  $\#\{Y \leq x\}$ . Note that the above quantities are basically rank quantities.

The first result is one by Smirnov.

**Theorem.** Given the statistic

$$D_{mn} = \sup_x |F_n(x) - F_m(x)|,$$

with  $X_1, \dots, X_m, Y_1, \dots, Y_n \sim F(x)$ . The distribution

$$\Phi_{mn}(\lambda) = \Pr \left( D_{mn} \leq \lambda \sqrt{\frac{mn}{m+n}} \right),$$

is independent of  $F(x)$ .

The above theorem states that the distribution under the null hypothesis for the Kolmogorov-Smirnov statistic is independent of  $F(x)$ .

We now sketch why this is true.

We first use the idea of symmetrization which we first encountered in the symmetrization proposition on page 16. For simplicity we assume that  $n = m$  in this context the symmetrization lemma on page 16 stated that: Given two draws from a distribution  $x_1, \dots, x_n$  and  $x'_1, \dots, x'_n$

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \frac{1}{n} \sum_{i=1}^n f(x'_i) \right| \geq \epsilon \right) \leq 2 \Pr \left( \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x) \right| \geq \epsilon/2 \right).$$

A result similar to the above was used to Kolmogorov and Smirnov to show that

$$\Pr \left( D_{mn} \leq \lambda \sqrt{\frac{mn}{m+n}} \right) \approx \Pr \left( D_n \leq \lambda \sqrt{2n} \right),$$

where

$$D_n = \sup_x |F_n(x) - F(x)|,$$

with  $X_1, \dots, X_n \sim F(x)$ .

In a paper that appeared in the Italian Journal of the Actuarial Institute in 1933 Kolmogorov proved the convergence of the empirical distribution function to the distribution function. This result was used by Smirnov in 1939 to derive the KS test result (Smirnov was a student of Kolmogorov).

Kolmogorov showed that:

**Theorem.** Given the statistic

$$D_n = \sup_x |F_n(x) - F(x)|,$$

with  $X_1, \dots, X_n, Y_1$ . The distribution

$$\Phi_n(\lambda) = \Pr \left( D_n \leq \lambda \sqrt{n} \right),$$

is independent of  $F(x)$ . In addition the limiting distribution is

$$\lim_{n \rightarrow \infty} \Phi_n(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 \lambda^2}.$$

A key idea in the proof of the result was the following lemma which is at the heart of the reason why rank statistics are nonparametric and independent of the distribution  $F(x)$ . The essence of this lemma is that looking at the difference between  $F_n(x) - F(x)$  is equivalent to looking at the difference between  $U_n(x) - U(x)$  where  $U(x)$  is the distribution function for the uniform distribution in the interval  $[0, 1]$  and  $U_n(x)$  is the empirical distribution function for  $n$  observations drawn iid from  $U[0, 1]$ .

**Lemma.** *The distribution  $\Phi_n(\lambda)$  is independent of  $F(x)$  if  $F(x)$  is continuous.*

*Proof.* Let  $X$  be a random variable with continuous distribution function  $F(X)$ , the random variable  $Y = F(X)$  has the following distribution  $F^{(0)}(x)$

$$\begin{aligned} F^{(0)}(x) &= 0, & x \leq 0; \\ F^{(0)}(x) &= x, & 0 \leq x \leq 1; \\ F^{(0)}(x) &= 0, & 1 \leq x. \end{aligned}$$

The above can be restated as  $Y$  is distributed as the uniform distribution on the interval  $[0, 1]$ . Given that  $F_n(x)$  and  $F_n^{(0)}(x)$  represent the empirical distribution functions for  $X$  and  $Y$  after  $n$  observations the following hold:

$$\begin{aligned} F_n(x) - F(x) &= F_n^{(0)}[F(x)] - F^{(0)}[F(x)], \\ &= F_n^{(0)}(y) - F^{(0)}(y) \\ \sup_x |F_n(x) - F(x)| &= \sup_x |F_n^{(0)}(y) - F^{(0)}(y)|. \quad \square \end{aligned}$$

The implication of the above lemma is that to study the distribution under the null hypothesis for the difference of distribution functions it suffices to study the uniform distribution.

The above lemma can be used to analyze the Mann-Whitney statistic since the difference in the average ranks of

$$\begin{aligned} A &: X_1, \dots, X_m \sim F_A \\ B &: Y_1, \dots, Y_n \sim F_B, \end{aligned}$$

can be written as

$$\bar{R}_A - \bar{R}_B = \frac{1}{m} \sum_{i=1}^m F_m(x_i) - \frac{1}{n} \sum_{i=1}^n F_n(x_i),$$

and the above lemma holds for this case as well. Note, the Mann-Whitney statistic can be rewritten in terms of the difference in average ranks of the two samples  $A$  and  $B$ .

Both the Mann-Whitney and KS statistics are nonparametric and so in the context of adaptability these are good tests. The general question of what is a good hypothesis test or is test A better than test B has still not been addressed. This question is typically addressed via the likelihood ratio testing framework and the Neyman-Pearson Lemma.

We start with the case of two simple hypotheses. Again these hypotheses are simple since the densities are completely specified under the null and alternative hypotheses.

$$\begin{aligned} H_0 &: X \sim p(X|H_0 = T) \\ H_A &: X \sim p(X|H_A = T). \end{aligned}$$

Given the sample  $X = \{X_1, \dots, X_n\}$  we can write down the likelihood ratio

$$\Lambda = \frac{p(X|H_0)}{p(X|H_A)}.$$

It would seem reasonable to reject  $H_0$  for small values of  $\Lambda$ .

**Lemma** (Neyman-Pearson). *Suppose the likelihood ratio test rejects  $H_0$  whenever  $\frac{p(X|H_0)}{p(X|H_A)} < c$  has significance level  $\alpha$*

$$\Pr\left(\frac{p(X|H_0)}{p(X|H_A)} < c\right) = \alpha.$$

*Then any other test which has significance level  $\alpha^* \leq \alpha$  has power less than or equal to the likelihood ratio test.*

This lemma address the issue of optimality for simple hypotheses.

**Example.** *We draw  $n$  observations iid from a normal distribution,*

$$X_1, \dots, X_n \sim N(\mu, \sigma^2).$$

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu = \mu_A. \end{aligned}$$

$$\Lambda = \frac{p(X|\mu_0, \sigma^2)}{p(X|\mu_A, \sigma^2)} = \frac{\exp(-\sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2)}{\exp(-\sum_{i=1}^n (x_i - \mu_A)^2 / 2\sigma^2)}.$$

$$\begin{aligned} \log(\Lambda) &= \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_A)^2, \\ &= \sum_i x_i^2 - 2\mu_0 \sum_i x_i + n\mu_0^2 - \sum_i x_i^2 + 2\mu_A \sum_i x_i - n\mu_A^2 \\ &= 2n\bar{X}(\mu_0 - \mu_A) + n\mu_A^2 - n\mu_0^2. \end{aligned}$$

*So we can use a statistic  $t(X)$*

$$t = 2n\bar{X}(\mu_0 - \mu_A) + n\mu_A^2 - n\mu_0^2,$$

*as our statistic to reject the null. Under the null hypothesis*

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right),$$

*so for the case where  $\mu_0 - \mu_A < 0$  the likelihood ratio test is a function of  $\bar{X}$  and is small when  $\bar{X}$  is small. In addition*

$$\Pr(\bar{X} \geq x_0) = \Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{x_0 - \mu_0}{\sigma/\sqrt{n}}\right),$$

so we can solve

$$\frac{x_0 - \mu_0}{\sigma/\sqrt{n}} = z(\alpha).$$

Which is the most powerful test under this model.

The problem with the likelihood ratio test framework is that in general the alternative hypothesis is not simple but composite. A typical situation would be

$$\begin{aligned} H_0 &: X \sim N(\mu_0, \sigma^2) \\ H_A &: X \sim N(\mu \neq \mu_0, \sigma^2), \end{aligned}$$

where the distribution under the alternative is not specified but forms a family of distributions. In this setting likelihood ration tests can be generalized to the concept of generalized likelihood ration tests which also have optimality conditions but these conditions are more subtle and we will not study these conditions.

Given the sample  $X = \{X_1, \dots, X_n\}$  drawn from a density  $p(X|\theta)$  with the hypotheses

$$\begin{aligned} H_0 &: \theta \in \omega_0 \\ H_A &: \theta \in \omega_A, \end{aligned}$$

where  $\omega_0 \cap \omega_A = \emptyset$  and  $\omega_0 \cup \omega_A = \Omega$ . The generalized likelihood ratio is defined as

$$\Lambda^* = \frac{\max_{\theta \in \omega_0} p(X|H_0(\theta))}{\max_{\theta \in \omega_A} p(X|H_A(\theta))}.$$

For technical reasons we will work with a slight variation of the above likelihood ration

$$\Lambda = \frac{\max_{\theta \in \omega_0} p(X|H_0(\theta))}{\max_{\theta \in \Omega} p(X|H_A(\theta))}.$$

Note that  $\Lambda = \min(\Lambda^*, 1)$  so small values of  $\Lambda^*$  correspond to small values of  $\Lambda$  so in the rejection region using either one is equivalent.

**Example.** We draw  $n$  observations iid from a normal distribution,

$$X_1, \dots, X_n \sim N(\mu, \sigma^2).$$

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0. \end{aligned}$$

The generalized likelihood is

$$\Lambda = \frac{\frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2)}{\max_{\mu} \frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2)},$$

the denominator is maximized at  $\mu = \bar{X}$ . So we can write

$$\begin{aligned} -2 \log \Lambda &= \frac{1}{\sigma^2} \left( \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{X})^2 \right) \\ &= n(\bar{X} - \mu_0)^2 / \sigma^2. \end{aligned}$$

We computed previously that if  $X_i \sim N(\mu_0, \sigma^2)$  that  $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ . The distribution of the square of a normal random variable is the chi-square distribution so

$$t = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \sim \chi_1^2,$$

where  $\chi_1^2$  is the chi-square distribution with one degree of freedom and we would reject the null hypothesis when

$$\frac{n(\bar{X} - \mu_0)^2}{\sigma^2} > \chi_1^2(\alpha).$$

We will apply the generalized likelihood ratio to two problems in classical genetics. However, the relevant statistical model involved in both problems requires the multinomial distribution which we now introduce.

There are  $n$  independent trials where for each trial one of  $r$  possibilities can occur each with probability  $p_1, \dots, p_r \geq 0$  where  $\sum_{i=1}^r p_i = 1$ . The actual counts from the  $n$  independent trials are  $n_1, \dots, n_r$  where  $\sum_{i=1}^r n_i = n$ . The (joint) distribution on the above counts is

$$\Pr(n_1, \dots, n_r) = \frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i^{n_i}.$$

We will parameterize the multinomial as  $M(p, n, r)$  where  $p = \{p_1, \dots, p_r\}$ .

Given observations  $X = \{n_1, \dots, n_r\}$  from the multinomial distribution with  $r$  possibilities and

$$\begin{aligned} H_0 &: X \sim M(\theta, n, r) \text{ with } \theta \in \omega_0 \\ H_A &: X \sim M(\theta, n, r) \text{ with } \theta \in \omega_A. \end{aligned}$$

We write down the likelihood ratio as

$$\Gamma = \frac{\max_{\theta \in \omega_0} \frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i(\theta)^{n_i}}{\max_{\theta \in \Omega} \frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i(\theta)^{n_i}}.$$

The maximization in the denominator is unconstrained so the unconstrained maximal likelihood estimate results in

$$\hat{p}_i = \frac{n_i}{n}.$$

The maximization in the numerator is constrained and we denote the estimate as

$$\hat{\theta} = \arg \max_{\theta \in \omega_0} \left[ \frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i(\theta)^{n_i} \right].$$

If we plug the above back into the likelihood ratio

$$\begin{aligned} \Gamma &= \frac{\frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i(\hat{\theta})^{n_i}}{\frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r \hat{p}_i^{n_i}} \\ &= \prod_{i=1}^r \left( \frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{n_i}, \\ &= \prod_{i=1}^r \left( \frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{n \hat{p}_i}, \end{aligned}$$

the last line comes from  $n_i = \hat{p}_i n$ . Taking the log

$$\begin{aligned} -2 \log \Gamma &= -2n \sum_{i=1}^r \hat{p}_i \log \left( \frac{p_i(\hat{\theta})}{\hat{p}_i} \right) \\ &= 2 \sum_{i=1}^r O_i \log \left( \frac{O_i}{E_i} \right), \end{aligned}$$

where  $O_i = n\hat{p}_i$  and  $E_i = np_i(\hat{\theta})$  are the observed and expected counts, respectively. If the dimensionality of the model space  $\omega_0$  is  $k$  then under the null hypothesis

$$t = 2 \sum_{i=1}^r O_i \log \left( \frac{O_i}{E_i} \right) \sim \chi_{r-k-1}^2,$$

where  $\chi_{r-k-1}^2$  is the chi-square distribution with  $r - k - 1$  degrees of freedom.

**Example.** Under Hardy-Weinberg equilibrium the genotypes  $AA$ ,  $Aa$ , and  $aa$  occur in a population with frequencies  $(1 - \tau)^2$ ,  $2\tau(1 - \tau)$ , and  $\tau^2$ . In a sample from the Chinese population in Hong Kong in 1937 blood types occurred with the frequencies given in the table below. Given the observed numbers and the Hardy-Weinberg equilibrium model we can estimate  $\tau$  using maximum likelihood,  $\hat{\tau} = .4247$ . This allows us to compute the expected counts under our model which is given in the table as well.

	$M$	$MN$	$N$
Observed	342	500	187
Expected	340.6	502.8	185.6

Given the above data our hypotheses are

$$H_0 : X \sim M(\{(1 - \tau^2), 2\tau(1 - \tau), \tau^2\}, n = 1029, r = 3)$$

$$H_A : X \sim M(\theta, n = 1029, r = 3) \text{ with } \theta \neq \{(1 - \tau^2), 2\tau(1 - \tau), \tau^2\},$$

so the null and alternate are both multinomial however the null assumes the Hardy-Weinberg model. For this case

$$-2 \log \Gamma = 2 \sum_{i=1}^3 O_i \log \left( \frac{O_i}{E_i} \right) = .032,$$

the likelihood is .98 and  $r - k - 1 = 1$  since  $r = 3$  and  $k = 1$  so the  $p$ -value is .86. There is no good reason to reject the Hardy-Weinberg model.

**Example** (Fisher's reexamination of Mendel's data). Mendel liked to cross peas. He crossed 556 smooth, yellow male peas with wrinkled, green female peas. According to the genetic theory he developed the frequency of the baby peas should be

Smooth yellow	$\frac{9}{16}$
Smooth green	$\frac{3}{16}$
Wrinkled yellow	$\frac{3}{16}$
Wrinkled green	$\frac{1}{16}$

The observed and expected counts are given in the following table.

Type	Observed count	Expected count
Smooth yellow	315	312.75
Smooth green	108	104.25
Wrinkled yellow	102	104.25
Wrinkled green	31	34.75

Given the above data our hypotheses are

$$H_0 : X \sim M\left(\left\{\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right\}, n = 556, r = 4\right)$$

$$H_A : X \sim M(\theta, n = 556, r = 4) \text{ with } \theta \neq \left\{\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right\}$$

so the null and alternate are both multinomial however the null assumes what became Mendel's law. For this case

$$-2 \log \Gamma = 2 \sum_{i=1}^4 O_i \log \left(\frac{O_i}{E_i}\right) = .618,$$

the likelihood is .73 and  $r - k - 1 = 3$  since  $r = 4$  and  $k = 1$  so the  $p$ -value is .9. There is no good reason to reject the Mendel's law.

Mendel did this experiment many times and Fisher pooled the results in the following way. Two independent experiments give chi-square statistics with  $p$  and  $r$  degrees of freedom respectively. Under the null hypothesis that the models were correct the sum of the test statistic would follow a chi-square distribution with  $p+r$  degrees of freedom. So Fisher added the chi-square statistics for all the independent experiments that Mendel conducted and found a  $p$ -value of .99996 !

### 3.0.12. Multiple hypothesis testing

In the various "omics" the issue of multiple hypothesis testing arises very often. We start with an example to illustrate the problem.

**Example.** The expression level of 12,000 genes can be measured with the technology available in one current platform. We measure the expression level of these 12,000 genes for 30 breast cancer patients of which  $C_1$  are those with ductal invasion and  $C_2$  are those with no ductal invasion. There are 17 patients in  $C_1$  and 13 in  $C_2$ . We consider a matrix  $x_{ij}$  with  $i = 1, \dots, 12,000$  indexing the genes and  $j = 1, \dots, 30$  indexing the patients. Assume that for each gene we use a  $t$ -test to determine if that gene is differentially expressed under the two conditions of for each  $i = 1, \dots, 12,000$  we assume  $X_{ij} \sim N(\mu, \sigma^2)$

$$H_0 : \mu_{C_1} = \mu_{C_2}$$

$$H_A : \mu_{C_1} \neq \mu_{C_2},$$

We set the significance level of each gene to  $\alpha = .01$  and we find that we reject the null hypothesis for 250 genes. At this point we need to stop and ask ourselves a basic question.

Assume  $H_0$  is true for  $i = 1, \dots, m = 12,000$  and we observe  $n = 30$  observations that are  $N(\mu, \sigma^2)$  split into groups of 13 and 17. We can ask about the distribution of the following two random variables

$$\xi = \# \text{ rejects at } \alpha = .01 \quad | \quad H_0 = T \quad \forall i = 1, \dots, m$$

$$\xi = \# \text{ times } t \geq t_{df}(.01) \quad | \quad H_0 = T \quad \forall i = 1, \dots, m$$

where  $t_{df}$  is the  $t$ -distribution with degree of freedom  $df$ . We can also ask whether  $\mathbb{E}\xi \approx 120$  or  $\mathbb{E}\xi \gg 120$ . This is the question addressed by multiple hypothesis testing correction.

The following contingency table of outcomes will be used ad nauseum in understanding multiple hypothesis testing.

	Accept null	Reject null	Total
Null true	$U$	$V$	$m_0$
Alternative true	$T$	$S$	$m_1$
	$W$	$R$	$m$

The main quantity that is controlled or worried about in multiple hypothesis testing is  $V$  the number of false positives or false discoveries. The other possible error is the type II error or the number of false negatives or missed discoveries. The two main ideas are the Family-wise error rate (FWER) and the False Discovery Rate (FDR). We start with the FWER.

3.0.12.1. *FWER*. The family-wise error rate consists of finding a cutoff level or  $\alpha$  value for the individual hypothesis tests such that we can control

$$\alpha_{FWER} = \Pr(V \geq 1),$$

this means we control the probability of obtaining any false positives. Our objective will be to find an  $\alpha$ -level or cutoff for the test statistic of the individual tests such that we can control  $\alpha_{FWER}$ .

The simplest correction to control for the FWER is called the Bonferroni correction. We derive this correction

$$\begin{aligned} \alpha_{FWER} &= \Pr(V \geq 1), \\ &= \Pr(\{T_1 > c\} \cup \{T_2 > c\}, \dots, \cup \{T_m > c\} | H_0) \\ &\leq \sum_{i=1}^m \Pr(\{T_i > c\}) \\ &\leq m \Pr(\{T > c\}) \\ &\leq m\alpha, \end{aligned}$$

so if we want to control  $\alpha_{FWER}$  at for example at .05 we can find a cutoff or select for  $\alpha = \frac{.05}{m}$  for the individual hypotheses. There is a huge problem with this correction the inequality between steps 2 and 3 can be large if the hypothesis are not disjoint (or the union bound sometimes sucks).

Let us make this concrete by using the example from the beginning of this section.

**Example.** *In the previous example we use a t-test to determine if a gene is differentially expressed under the two conditions, for each  $i = 1, \dots, 12,000$  we assume  $X_{ij} \sim N(\mu, \sigma^2)$*

$$\begin{aligned} H_0 &: \mu_{c_1} = \mu_{c_2} \\ H_A &: \mu_{c_1} \neq \mu_{c_2}, \end{aligned}$$

*We set the significance level of each gene to  $\alpha = .01$  this gives us the Bonferroni correction of  $\alpha_{FWER} = .01 \times 12,000 = 120$ , which is a joke. In addition, the assumption that the hypotheses, genes, are independent is ludicrous.*

We now use an alternative approach based upon a computational approach that falls under the class of permutation procedures. This lets us avoid the union

bound and control the FWER more accurately. We will also deal with the issue that t-distribution assumes normality and our data may not be normal. We will develop this procedure in the context of the previous example.

**Example.** We will use a t-statistic to determine if a gene is differentially expressed under the two conditions. However, for each  $j = 1, \dots, 12,000$  we do not make distributional assumptions about the two distributions

$$\begin{aligned} H_0 &: C_1 \text{ and } C_2 \text{ are exchangeable} \\ H_A &: C_1 \text{ and } C_2 \text{ are not exchangeable,} \end{aligned}$$

by exchangeable we mean loosely  $\Pr(x, y | y \in C_1) = \Pr(x, y | y \in C_2)$ .

For each gene we can compute the t-statistic:  $t_i$ . We then repeat the following procedure  $\pi = 1, \dots, \Pi$  times for each gene

- (1) permute labels
- (2) compute  $t_i^\pi$ .

For each gene  $i$  we can get a p-value by looking at where  $t_i$  falls in the ecdf generated from the sequence  $\{t_i^\pi\}_{\pi=1}^\Pi$ ,

$$p_i = \hat{\Pr}(\xi > t_i | \{t_i^\pi\}_{\pi=1}^\Pi).$$

The above procedure gives us a p-value for the individual genes without an assumption of normality but how does it help regarding the FWER.

The following observation provides us with the key idea

$$\begin{aligned} \alpha_{FWER} &= \Pr(V \geq 1), \\ &= \Pr(\{T_1 > c\} \cup \{T_2 > c\}, \dots, \cup \{T_m > c\} | H_0) \\ &= \Pr\left(\max_{i=1, \dots, m} \{T_i > c\} | H_0\right). \end{aligned}$$

This suggests the following permutation procedure For each gene we can compute the t-statistic:  $\{t_i\}_{i=1}^m$ , where  $m = 12,000$ . Then repeat the following procedure  $\pi = 1, \dots, \Pi$

- (1) permute labels
- (2) compute  $t_i^\pi$  for each gene
- (3) compute  $t^\pi = \max_{i=1, \dots, m} t_i^\pi$ ,

for each gene we can compute the FWER as

$$p_{i,FWER} = \hat{\Pr}(\xi > t_i | \{t^\pi\}_{\pi=1}^\Pi).$$

One very important aspect of the permutation procedure is we did not assume that the genes were independent and in some sense modeled the dependencies.

For many problems even this approach with a more accurate estimation of the FWER p-value does not give us significant hypotheses because the quantity we are trying to control  $\Pr(V \geq 1)$  is too stringent.

3.0.12.2. *FDR.* The false discovery rate consists of finding a cutoff level or  $\alpha$  value for the individual hypothesis tests such that we can control

$$q_{FDR} = \mathbb{E} \left[ \frac{V}{R} \right],$$

this is controlling the proportion of false positives among the hypotheses we reject. For the reason that the above is not well defined when  $R = 0$  we adjust the statistic so that we condition on there being rejects

$$q_{pFDR} = \mathbb{E} \left[ \frac{V}{R} | R > 0 \right].$$

We first illustrate the procedure with the permutation procedure we introduced previously as an example and then we will look at the more classical parametric case.

**Example.** *We will use a  $t$ -statistic to determine if a gene is differentially expressed under the two conditions. However, for each  $j = 1, \dots, 12,000$  we do not make distributional assumptions about the two distributions*

$$H_0 : C_1 \text{ and } C_2 \text{ are exchangeable}$$

$$H_A : C_1 \text{ and } C_2 \text{ are not exchangeable,}$$

by exchangeable we mean loosely  $\Pr(x, y | y \in C_1) = \Pr(x, y | y \in C_2)$ .

For each gene we can compute the  $t$ -statistic:  $\{t_i\}_{i=1}^m$ , where  $m = 12,000$ . Then repeat the following procedure  $\pi = 1, \dots, \Pi$

- (1) permute labels
- (2) compute  $t_i^\pi$  for each gene

note that the statistics  $\{t_i^\pi\}_{i,\pi}$  were all drawn under the assumption that the null is true. So if we reject any of them they would be a false positive. The statistics  $\{t_i\}_{i=1}^m$ , are drawn from a combination of hypotheses for which the null hypothesis holds true and those for which the alternative is true. Define the ranked list of the statistics under the null hypothesis as

$$\{\text{Null}_{(i)}\}_{i=1}^{\Pi \times m} = \text{ranked}\{t_i^\pi\} \quad \text{for } i = 1, \dots, m, \pi = 1, \dots, \Pi.$$

Similarly we can define the ranked list of the statistics coming from the set of the alternative and the null as

$$\{\text{Tot}_{(i)}\}_{i=1}^m = \text{ranked}\{t_i\} \quad \text{for } i = 1, \dots, m.$$

We now look at a series of cutoffs corresponding to

$$c = \text{Tot}_{(1)}, \text{Tot}_{(2)}, \dots, \text{Tot}_{(k)}.$$

For each value of  $c$  we can compute the following two quantities

$$\begin{aligned} \%R(c) &= \frac{\#\{\text{Null}_{(i)}\} \leq c}{m \times \Pi}, \\ \%V(c) &= \frac{\#\{\text{Tot}_{(i)}\} \leq c}{m}, \end{aligned}$$

from which we can compute the  $pFDR$  for the given cutoff  $c$

$$q_{pFDR}(c) = \frac{\%V(c)}{\%R(c)}.$$