PhD Oral Exam Notes **Fall 2016**

Probability and Statistics

Ye Li *Scribes: 520.651: Random Signal Analysis*

This notes cover basic knowledge in probability theory and statistics in Electrical Engineering and Image Science.

1 Introduction to Probability

1.1 Probability space

A probability space is a mathematical construct that models a real-world process (like tossing a coin OR tossing two coins) consisting of states (H , T OR HH,TT,HT,TH) that occur randomly. A probability space is usually constructed with a specific kind of situation or experiment in mind.

A probability space consists of three entities:

1. A **sample space** is a set of all possible outcomes (not combination of outcomes) of the random experiment, which usually is denoted by Ω.
   Example:

$$\omega \in \Omega$$

   where ω is an experimental outcome. For example, if the experiment is tossing a coin, all possible outcomes are just $\{head, tail\}$. This set is called **sample space**. For tossing four coins, the sample space is

$$
\begin{gathered}
HHHH \\
HHHT \\
HHTT \\
HHTH \\
HTTT \\
HTHH \\
HTHT \\
HTTH \\
TTTT \\
TTHH \\
TTHT \\
TTTH \\
THTT \\
THHH \\
THHT \\
THTH
\end{gathered}
\tag{1}
$$

2. An **event space** is customized to the way you define the random variable. An event space consists a set of events (one event can be you get exactly zero head before 1st tail, another can be you get 1 head before the first tail; an event corresponds to one value that the random variable can take) and is usually denoted by F, where each event is a set containing zero or more outcomes. As an example, we can define a random variable that records the number of heads before the first tail on the sample space specified above. The random variable takes 4 possible values, namely 0, 1, 2, and 3. One event that corresponds to 0, which describes that you get 0 head before the 1st tail, would thus consist of the following outcomes:

$$
\begin{gathered}
HHHH \to 0 \\
TTTT \to 0 \\
TTHH \to 0 \\
TTHT \to 0 \\
TTTH \to 0 \\
THTT \to 0 \\
THHH \to 0 \\
THHT \to 0 \\
THTH \to 0
\end{gathered}
\tag{2}
$$

The **event space** is a space that consists of all events (in the example 4 events) and therefore represents all possible combinations of outcomes from the experiment.

Example:

$$\omega \in A \in \Omega, F = \{A, B, ..\}$$

where $A = \{1, 3, 5\}$ and $\Omega = \{1, 2, 3, 4, 5, 6\}$. ω is a point or an atom of Ω, i,e., $\omega = 1$. Here A represents an event of only getting odd numbers.

3. Given an experiment with sample space Ω, a **probability measure** is a function, denoted by P, defined on the space of all events of the experiments and taking values in the interval $[0, 1]$. For every event A, $P(A)$ provides a numerical assessment of the likelihood that the event A occurs; the quantity $P(A)$ is the probability of A. For any event E,

$$0 \geq P(E) \leq 1$$

For any sequence of events $E_1, E_2, ...$ which are mutually disjoint,

$$P(\cup_{n=1} E_n) = \sum_{n=1} P(E_n)$$

P is a function mapping event in event space to a probabilistic number. Probabilities are defined on event space, which is a collection of subsets of Ω, not Ω

The standard representation of a **probability space** is a triple with the above three entities in their respective order, i.e.,

$$(\Omega, F, P)$$

where the pair (Ω, F) is referred as a **measurable space**, which describes the outcomes and modes of observation of the experiment without reference to the likelihood of the observables. In theory, the same **measureable space** can give rise to many different probability space.

Brain Teasers:

1. How many subsets(possible events) are in $\Omega = \{H, T\}$?

$$\{H\}, \{T\}, \{H, T\}, \phi$$

We say that the event E occurs if the outcome of the experiment is contained in E (recall an event is a set). So, we may think this as an event of getting any outcome - as long as you toss the coin you'll get an outcome, thus this event has probability 1 or, $P(\Omega) = 1$. ϕ can be thought as an event of getting nothing.

2. Definition of a power set: The power set of any set S is the set of all subsets of S, including the empty set and S itself. If S is the set $\{x, y, z\}$, then the subset of S are:

$$\{\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}$$

and hence the power set of S is $\{\{\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$

If we think S as our sample space in this case which contains 3 possible outcomes of an experiment, then as shown above, we can have as many as 8 events in the event space. Remember, we construct our event space based on the definition of the random variable in context. Here, if we specify a random variable to denote whether or not the outcome contains z. Then, the event $\{x, y\}$ above can be thought as an event that you do not get z in the outcome and for example assigned a value 0. And, the other event in the event space is the event that you get z in the outcome and assigned a value 1.

But the richer the event space, the more work is required for the probability measure function P. So it's better to have a minimum/lean event space. That is try to have less events in your event space, or equivalently make your random variable equal to as few real values as possible. Because the more real values your random variable can take, the more work is required later to define probability (likelihood of an event). Say, if you have a random variable that takes on 4 different real values, then in order to construct its probability distribution function, you would need to count the number of occurrence (frequency) for each of the values that the random variable can take.

Note: It's not necessary that a point ω should be of the same form as the outcome of the physical experiment; it merely suggests that one should be able to set up a correspondence between the actual outcomes and the points in Ω. For example, the sample space $\Omega = (0, 1]$ can adequately represent the outcomes of a sequence of coin tosses, despite of the fact that the points in Ω are not themselves binary sequences. This is because it is possible to identify every ω with a distinct binary sequence by taking its binary expansion (conversely, every binary sequence that does not converge to 0 can be identified with a distinct point in $(0, 1]$).

1.2 Field

Motivation: Why need field?
If we're talking about event A and B,we want to talk about their probabilities and the probability of $A \bigcup B$. Since probabilities are defined on events, in order for that to happen, $A \bigcup B$ needs to be an event or event collection. Below are criteria that a field must meet:
An event space F becomes a field if:

1. $\phi, \Omega \in F$ (The empty set ϕ is called the impossible event. Its probability is always zero. The entire sample space Ω is called the sure event. Its probability is always one, i.e., $P(\Omega) = 1$)
2. if $A \in F => A^c \in F$ (closure under complementation)
3. if $A, B \in F => A \cup B \in F$ (closure under union)
4. $A \in F, B \in F => A \cap B \in F$ (closure under intersection)
5. $A_1, ..., A_n \in F => A_1 \cup ... \cup A_n \in F$ (closure under finite unions)

Example(two extremes):

1. Two elements $\{\phi, \Omega\}$ make the smallest field; a totally valid field.
2. 2^Ω, where 2^Ω is defined as the power set (the collection of all subsets of Ω)

1.3 Event spaces and sigma-fields

An event space, F, consists of events, namely possible combinations of outcomes from the experiment and it can be defined as follows:

$$F = \{A = \bigcup_{i=I}^{\infty} C_i, I \in \{1, 2, 3, 4, .., k\}\}$$

where $\Omega = \bigcup_{i=I}^{\infty} C_i$, C_i is an event, like a slice in a pizza. k is like the number of slices in the pizza. If $k = 2$, we will have a field like this: $(\Omega, \{\phi, A, A^c, \Omega\})$

Now we want to talk about the probability of each individual event as well as their union. For example A_i can be an event in the field but countable unions, namely, $\bigcup\limits_{i=I}^{\infty} A_i$ does not necessarily belong to the field. For example, if we define an event, A_i using a semi-open interval, $(0, \frac{1}{2} - \frac{1}{3i}]$. Then, its countable union, $\bigcup\limits_{i=I}^{\infty} A_i = (0, \frac{1}{2})$ will be outside of F because it cannot be expressed as a finite union of semi-open intervals hence violate (3). So for scenarios like this, we can't talk about its probability. Now we are going to modify the axioms of a field to below:

1. $\phi, \Omega \in F$ (The empty set ϕ is called the impossible event. Its probability is always zero. The entire sample space Ω is called the sure event. Its probability is always one, i.e., $P(\Omega) = 1$)
2. if $A \in F => A^c \in F$ (closure under complementation)
3. $A_1, ..., A_n \in F => \bigcup\limits_{i=I}^{\infty} A_i \in F$ (closure under countable instead of finite unions)

And, these axioms make the event space F suitable for a σ field (a field that's closed under countable unions). For example, $F = \{\phi, \Omega\}$ makes (Ω, F) the smallest σ field and $F = 2^\Omega$ makes F the largest σ field. Also, the number of elements in F is 2^n. When $n = 2$, there will be 4 elements in F, i.e., $F = \{\phi, \Omega, A, A^c\}$

1.4 Generated (man-made) sigma-fields

If we were to generate a *sigma* field, we would want a minimal *sigma* field as we would like to keep the event space as lean as possible. Support $\Omega = (0, 1]$, and define the collection $\mathcal{G}$ by

$$\mathcal{G} = \{(0, 1/3], (2/3, 1]\}$$

To construct the minimum σ field, we need the following:

1. 1) ϕ and 2) $(0, 1]$ must lie in every σ field containing $\mathcal{G}$, as do 3) $(0, 1/3]$ and 4) $(2/3, 1]$ and 5) their union $(0, 1/3] \cup (2/3, 1]$, and so dose 6) its complement $(1/3, 2/3]$
2. By closure under union (closure to the events in $\mathcal{G}$), 7) $(0, 2/3]$ and 8) $(1/3, 1]$ must also lie in every σ field that containing $\mathcal{G}$

So the smallest σ field containing $\mathcal{G}$ is the following:

$$\mathcal{F} = \{\phi, \Omega, (0, 1/3], (1/3, 2/3], (2/3, 1](0, 2/3], (1/3, 1], (0, 1/3] \cup (2/3, 1]\}$$

1.5 The Borel field

As we had proved above that the field below is not a σ field.

$$\mathcal{F} = \{\phi\} \cup \{A : A = \bigcup_{i=I}^{\infty} (a_i, b_i], M < \infty, (a_i, b_i] \in (0, 1]\}$$

as countable unions such as $(0, a)$ is not in the field, which violates the closure under countable union criteria of a σ field. By including the countable union(s) to the field, $\mathcal{F}$ can be augmented to a minimal σ field, $\sigma(\mathcal{F})$. This σ field is called the Borel field of the unit interval, and is denoted by $\mathcal{B}((0, 1])$. The Borel field contains $\mathcal{F}$ and the countable unions such as $(0, a)$.

1.6 Random Variable

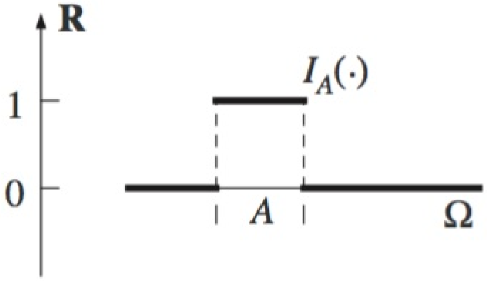
A real-valued function $\omega : X(\omega)$ defined for points ω in a sample space Ω is called a **random variable**. A random variable(r.v.) on a measurable space $(\Omega, \mathcal{F})$ is a real-valued function $X = X(.)$ on Ω such that for all $a \in R$, the set

$$X^{-1}(-\infty, a] = \{\omega : X(\omega) \leq a\}$$

lies in $\mathcal{F}$, i.e., the set is an event. For example, a special case of a r.v. is an indicator function $I_A(\omega)$. The indicator function is defined on the sample space Ω) for each experimental outcome, ω and maps each outcome to value of 1 and 0 (ONLY). But in general the values can be any real number.

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \in A^c \end{cases} \tag{3}$$

Graphically, we have



The set X^{-1} is given by

$$X^{-1}(-\infty, a] = \begin{cases} \phi, & \text{if } a < 0 \\ A^c, & \text{if } 0 \leq a < 1 \\ A, & \text{if } a = 1 \\ \Omega, & \text{if } a \geq 1 \end{cases} \tag{4}$$

Basically, a random variable's input is an experimental outcome and outputs a scalar. But note that probabilities are defined on event space which contains subsets of Ω
A less abstract example that combines sample space, random variable, and its probability density function is as follows:

Example An experiment consists of throwing a fair coin 4 times, you are asked to find the frequency function and the cumulative distribution function of the following random variable:
a) the number of heads before the first tail
b) the number of heads following the first tail
c) the number of heads minus the number of tails
d) the number of tails times the number of heads

Soln: In order to find the pdf and cdf of these random variables, we first need to list the sample space of the underlying experiment.

$$\begin{aligned}\Omega = \{&HHHH, HHHT, HHTT, HHTH, HTTT, HTHH, HTHT,\\ &HTTH, TTTT, TTHH, TTHT, TTTH, THHH, THHT, THTH\}\end{aligned} \tag{5}$$

For the random variable in a), we have

$$\begin{aligned}HHHH &\to 0\\ HHHT &\to 3\\ HHTT &\to 2\\ HHTH &\to 2\\ HTTT &\to 1\\ HTHH &\to 1\\ HTHT &\to 1\\ HTTH &\to 1\\ TTTT &\to 0\\ TTHH &\to 0\\ TTHT &\to 0\\ TTTH &\to 0\\ THTT &\to 0\\ THHH &\to 0\\ THHT &\to 0\\ THTH &\to 0\end{aligned} \tag{6}$$

Each arrow above is actually representing the functionality (what a r.v. does) of a random variable, which is a function that transfers experimental outcome in event space (automatically in sample space) to a real-valued number. After counting we have 9 zeros, 2 twos, 1 three, and 4 ones. So, for pdf we have

$$\begin{aligned}P(X=0) &= \frac{9}{16}\\ P(X=2) &= \frac{2}{16}\\ P(X=3) &= \frac{1}{16}\\ P(X=1) &= \frac{4}{16}\end{aligned} \tag{7}$$

These probabilities are also *probability measures*, which is a function that's defined on the event space, and takes values in $[0, 1]$ to provide numerical assessment of the likelihood that event A occurs, i.e., $P(A) = P(X = 2)$. And event A's event space is $A = \{HHTT, HHTH\}$. One event usually corresponds to one real-valued number which the random variable would actually take. In other words, probability frequency function of a random variable is actually composed of probability measures of events as different events get mapped to different real-valued numbers by the random variable.

1.7 Joint Distributions

Joint probability is 2 or more random variables defined on the same sample space. For example, for an experiment of a coin tossed twice, the sample space is

$$\Omega = \{HT, TH, TT, HH\}$$

Now define random variables on this sample space:

- Let random variable X be the vent that you get head on the first toss
- Let random variable Y be the total number of head in the trail (2 tosses in total)

Y/X	Yes	No
0	0	0.25
1	0.25	0.25
2	0.25	0

e.g., $P(X = Y, Y = 1) = 0.25$

1.8 Independence

If event A and B satisfy $P(A|B) = P(A|B^c)$ we say that A **does not depend** on B. The event A and B are called *independent* if and only if they satisfy

$$P(A \cap B) = P(A)P(B) \tag{8}$$

Two sets A and B are called **disjoint** if $A \cap B = \phi$. Two sets being disjoint has nothing to do with their probabilities and is purely a relation in the event space. But the notion of independence does depend on P. The definition of independence implies

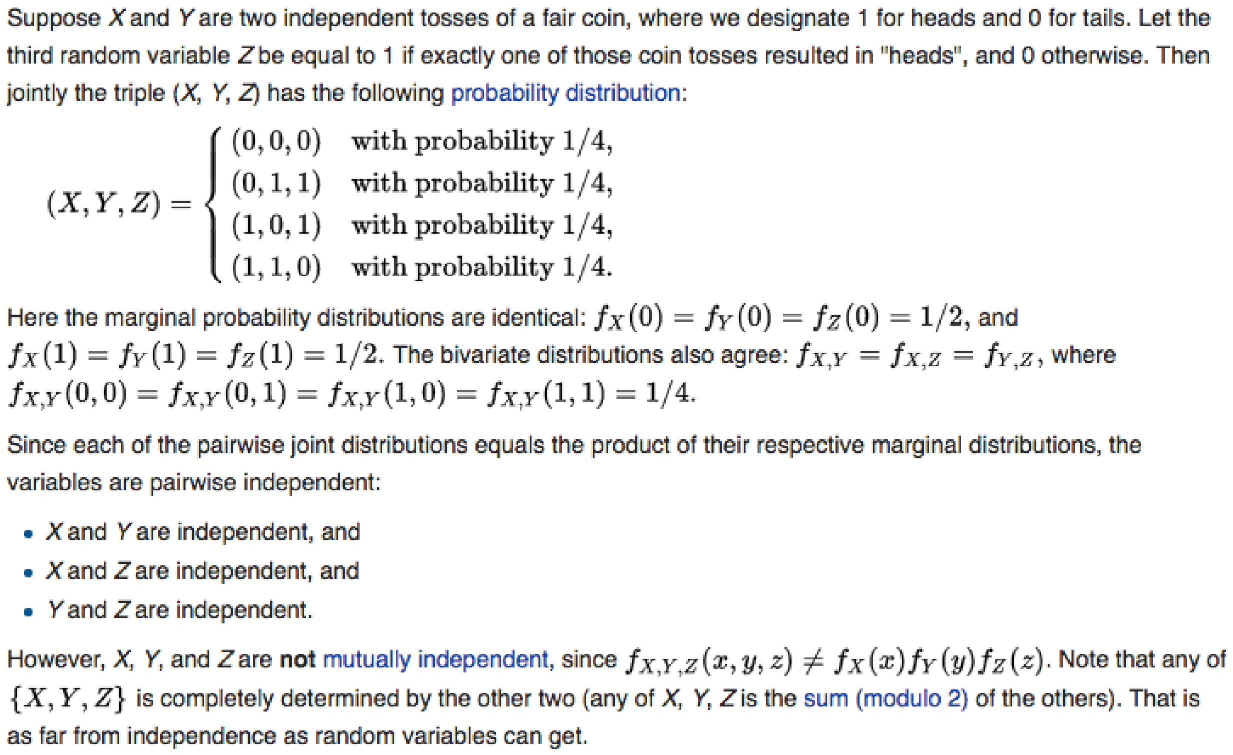
$$P(A|B) = P(A) \tag{9}$$

Also, if A and B are independent then so are A and B^c, A^c and B, A^c and B^c. If follows that if any one of the four pairs is independent then so are the other three. Again, the notion of independence is defined on probability and has nothing to do with the events being disjoint or not. Also, if $P(B) = 0$ or $P(B) = 1$ then A and B are independent as B will or will not happen for sure and thus has no effect on A.

A sequence of events A_j, j=1,2,... (now imagine that there are more than two events in the event space) are **mutually independent** if for every finite subset J containing two or more positive integers,

$$P(\cap_{j \in J} A_j) = \prod_{j \in J} P(A_j) \tag{10}$$

Note: **Pairwise independence** means only two events are independent of each other and does not imply mutual independence, as shown by the following example from Wikipedia:



Suppose X and Y are two independent tosses of a fair coin, where we designate 1 for heads and 0 for tails. Let the third random variable Z be equal to 1 if exactly one of those coin tosses resulted in "heads", and 0 otherwise. Then jointly the triple (X, Y, Z) has the following probability distribution:

$$(X,Y,Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probability } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{cases}$$

Here the marginal probability distributions are identical: $f_X(0) = f_Y(0) = f_Z(0) = 1/2$, and $f_X(1) = f_Y(1) = f_Z(1) = 1/2$. The bivariate distributions also agree: $f_{X,Y} = f_{X,Z} = f_{Y,Z}$, where $f_{X,Y}(0,0) = f_{X,Y}(0,1) = f_{X,Y}(1,0) = f_{X,Y}(1,1) = 1/4$.

Since each of the pairwise joint distributions equals the product of their respective marginal distributions, the variables are pairwise independent:

- X and Y are independent, and
- X and Z are independent, and
- Y and Z are independent.

However, X, Y, and Z are **not** mutually independent, since $f_{X,Y,Z}(x,y,z) \neq f_X(x)f_Y(y)f_Z(z)$. Note that any of $\{X,Y,Z\}$ is completely determined by the other two (any of X, Y, Z is the sum (modulo 2) of the others). That is as far from independence as random variables can get.

1.9 Conditional Probability

Given two events A and B the conditional probability is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{11}$$

The definition gives us the following relation

$$P(B|A) = \frac{P(A \cap B)P(B)}{P(A)} \tag{12}$$

and the law of total probability

$$P(A) = P(A \cap B)P(B) + P(A \cap B^c)P(B^c) \tag{13}$$

Combining this with the previous results, we get the **Bayes rule**

$$P(B|A) = \frac{P(A \cap B)P(B)}{P(A \cap B)P(B) + P(A \cap B^c)P(B^c)} \tag{14}$$

Also, note that the conditional probability is a probability as a function of its first argument, we can write $P(A^c|B) = 1 - P(A|B)$

Assume that we observe an event A in an experiment and B_n is a sequence of pairwise disjoint events that we can't observer but would like to make inference on, namely we want to know $P(B_k|A)$ - an entire distribution which has the same number of distinct values that A can take on. Before making any observations we know the prior probabilities $P(B_n)$ and we know the conditional probabilities $P(A|B_n)$. After we observer A, we compute the posterior probabilities $P(B_k|A)$ for each k (notice that there is a change of suffix here because now we are after the whole distribution not only the probabilities of the n discrete points in our prior distribution). To do so we generalize the law of total probability. Assume that $\sum_n P(B_n) - 1$. Then law of total probability is

$$P(A) = \sum_n P(A|B_n)P(B_N) \tag{15}$$

and the general Bayes rule gives,

$$P(B_k|A) = \frac{P(A|B_n)P(B_n)}{\sum_n P(A|B_n)P(B_n)} \tag{16}$$

1.10 Combinatorics and Probability

Permutation: Permutation differs from combination in that it cares about the order of the set. Combinations are selections of some members of a set where order is disregarded. For example, written as tuples, there are six permutations of the set $\{1, 2, 3, \}$, namely: (1,2,3), (1,3,2,), (2,1,3),(2,3,1), (3,1,2) and (3,2,1). These are all the possible orderings of this three-element set.

Combination: Combination is a way of selecting items from a collection, such that the order of selection does not matter. For example, given three fruits, say an apple, an orange and a pear, there are three combinations of two that can be drawn from the set: an apple, a pear; an apple and an orange; or a pear and an orange.

In combinatorics, there are fours kinds of counting problems:

- ordered sampling with replacement: For k finite sets $A_1, A_2, ..., A_k$ (each set has $|A_i|$ members), there are $|A_1| * |A_2| * ...|A_k|$ k-tuples of the form $(a_1, a_2, ..., a_k)$ where each $a_i \in A_i$
- ordered sampling without replacement (permutation): Given a set (only one set) A with cardinatlity $|A| = n$, there are $\frac{n!}{(n-k)!}$ k-tuples of the form $(a_1, a_2, ..., a_k)$ with distinct entries a_i, $a_i \in A$.
- unordered sampling without replacement (combination): If the ordering of the elements a_i is not important then there are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ (n choose k) k-tuples of the form $(a_1, a_2, ..., a_k)$ with distinct entries a_i. This differs from 2 by $k!$ in the denominator which accounts for the number of permutations that give the same combination when the order is ignored. For example,$k! = 2! = 2$ accounts the fact that $(1, 2)$ and $(2, 1)$ are equal. $k! = 3! = 6$ would account for the fact that (1,2,3), (1,3,2,), (2,1,3),(2,3,1), (3,1,2) and (3,2,1) are equal. For example, the string, "BUTTHEADEDD", we would have 11! of ways to sequence it. But the number of combinations should be equal to $\frac{11!}{2!2!3!}$. The denominator accounts for permutations of the letter "T", "E" and "D". So in general, the number of such sets possible is given by the **multinomial coefficient**

  $$\binom{n}{k_0, k_1, k_2, ..., k_{m-1}} = \frac{n!}{k_0!k_1!..k_{m-1}!} \tag{17}$$

  where $k_0 + k_1 + k_2 +, ..., +k_{m-1} = n$

  Binomial is just a special case of multinomial: $\binom{n}{n1, n_2} = \binom{n}{n_1} = \binom{n}{n_2}$ since $n_1 + n_2 = n$

- unordered sampling with replacement: Given a set A with $|A| = n$, there are $\binom{k+n-1}{k}$ ways in which different sets of length k can be chosen with replacement from n items.

For example, for a set of size n and a sample of size r, there are $n \times n \times n... = n^r$ different ordered samples with replacement and $n(n-1)(n-2)...(n-r+1)$ different ordered samples without replacement. Also, the number of ordered samples is equal to the number of unordered samples times the number of ways to order each sample($r!$). So the number of unordered samples is equal to $n(n-1)(n-2)...(n-r+1)/r!$, which is equal to $\frac{n!}{(n-r)!r!} = \binom{n}{r}$.

2 Discrete Random Variables

2.1 Probability mass functions

The **probability mass function (pmf)** of a discrete random variable X taking distinct values x_i is defined by

$$p_X(x_i) = P(X = x_i) \tag{18}$$

The Poisson random variable is used to model many different physical phenomena ranging from the photoelectric effect and radioactive decay to computer message traffic arriving at a queue for transmission. A random variable X is said to have a Poisson probability mass function with parameter $\lambda > 0$, denoted by $X \sim Poisson(\lambda)$ if ,

$$p_X(x_i) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2... \tag{19}$$

The **joint probability mass function** of X and Y is defined by

$$p_{XY}(x_i, y_i) = P(X = x_i, Y = y_i) \tag{20}$$

The **marginal probability mass functions** are given as

$$\begin{aligned} p_X(x_k) &= \sum_j p_{XY}(x_k, y_j) \\ p_Y(y_j) &= \sum_k p_{XY}(x_k, y_j) \end{aligned} \tag{21}$$

A pair of discrete random variables are **independent** if and only if their joint pmf factors into the product of their marginal pmfs:

$$p_{XY}(x_i, y_i) = p_X(x_i) p_Y(y_i) \tag{22}$$

2.2 Expectation

The **expectation, mean or average** of any discrete random variable X is defined by

$$E[X] = \sum_i x_i P(X = x_i) \tag{23}$$

or using the pmf notation,

$$E[X] = \sum_i x_i p_X(x_i) \tag{24}$$

Given a random variable X and let $Z = g(X)$ be a new random variable where $g(x)$ is a real-valued function. The expected value of Z is given by the $lawoftheunconsciousstatistacian(LOTUS)$.

$$E[Z] = E[g(X)] = \sum_i g(x_i) p_X(x_i) \tag{25}$$

In general if $g(x, y)$ is a real-valued function of two variables x and y, then

$$E[g(X, Y)] = \sum_i \sum_j g(x_i, y_i) p_{XY}(x_i, y_i) \tag{26}$$

The expectation operator is **linear**

$$E[aX + bY] = E[aX] + E[bY] = aE[X] + bE[Y] \tag{27}$$

If X and Y are independent then for any function $h(x)$ and $k(y)$

$$E[h(X)k(Y)] = E[h(X)]E[k(Y)] \tag{28}$$

The **correlation** between uncorrelated random variables is zero if and only if at least one of them has zero mean. If not then they are uncorrelated if and only if

$$E[(X - m_X)(Y - m_Y)] = 0 \tag{29}$$

If X and Y are independent then they are uncorrelated. The notion of being uncorrelated is weaker than that of **independence**.

2.3 Moments

The n^th moment, $n \geq 1$ of a real-valued random variable X is defined to be $E[X^n]$. The first moment of X is its mean, $m = E[X]$. The central moment of X is defined as $E[(X - m)^n]$. The variance is defined as the average squared deviation of X about its mean. It is the second central moment

$$var(X) = E[(X - m)^2] \tag{30}$$

The **variance** characterizes how likely it is to observe values of X far from its mean. When a random variable dose not have zero mean, it is often convenient to use the variance formula,

$$var(X) = E[X^2] - (E[X])^2 \tag{31}$$

The **standard deviation** of X is defined to be the positive square root of the variance. If X_1, X_2, ... be a sequence of uncorrelated random variables; i.e., for $i \neq j$, X_i and X_j are uncorrelated then the variance of the sum is the sum of variances,

$$var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) \tag{32}$$

2.4 The Markov and Chebyshev Inequalities

The Markov and Chebyshev Inequalities provide bounds on probabilities in terms of expectations that are more readily computable.

- The **Markov inequality** says that if X is a non-negative random variable, then for any $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a} \tag{33}$$

- The **Chebyshev inequality** says that for any random variable X and any $a > 0$,

$$P(|X| \geq a) \leq \frac{E[X^2]}{a} \tag{34}$$

- The following are special cases of the Chebyshev inequality

$$P(|X - \mu| \geq a) \leq \frac{var(X)}{a} \tag{35}$$

where $\mu = E[X]$, if $\sigma^2 = var(X)$, then taking $a = k\sigma$ yields

$$P(|X - \mu| \geq a) \leq \frac{1}{k^2} \tag{36}$$

2.5 Probability Generating Functions

Let X be a discrete random variable taking only non-negative integer values. The **probability generating function** of X is

$$G_X(z) = E[x^X] = \sum_{n=0}^{\infty} z^n P(X = n) \tag{37}$$

The probability generating function (pgf) is used to compute mean and variance and the probability mass function. The pgf of a sum of *independent* random variables is the product of the individual pgfs, i.e., if $Y = X_1 + X_2 + ... + X_n$ then

$$G_Y(z) = G_{X_1}(z) * G_{X_2}(z)...G_{X_n}(z) \tag{38}$$

This is called the **factorization property**. The probability generating function can be expanded as the series,

$$G_X(z) = P(X = 0) + zP(X = 1) + z^2 P(X = 2) + ... \tag{39}$$

Two different pmfs can't have the same pgf since pmf can be recovered from pgf as follows:

$$\frac{G_X^{(k)}(z)|_z = 0}{k!} = P(X = k) \tag{40}$$

The probability generating function can also be used to find moments. The K^th **factorial moment** of X is given as

$$G_X^{(k)}(z)|_z = 1 = E[X(X-1)(X-2)...(X-[k-1])] \tag{41}$$

where $G_X^{(k)}(z)$ is the k^th derivative of the probability generating function. A special case for $k = 1$ yields $G_X^{'}(1) = E[X]$

2.6 The Binomial Random Variable

In many problems, the key quantity of interest can be expressed in the form $Y = X_1 + ... + X_n$ where X_i are i.i.d. Bernoulli(p) random variables. For Bernoulli(p) random variable, $G_{X_i} = E[z^{X_i}] = z^0(1-p) + z^1 p = (1-p) + pz$ and so $G_Y(z) = [(1-p) + pz]^n$. Recovering pmf using the relation (1.35) gives

$$P(Y = k) = \frac{n!}{k!(n-k)!} p^k (1 - p^{n-k}) \tag{42}$$

Since $G_Y(z)$ is a polynomial of degree n, $G_Y^{(k)}(z) = 0$ and so $P(Y = k) = 0$, for all $k > n$. This random variable is called a **binomial(n,p) random variable. It counts how many times an event has occurred**. The binomial(n,p) random variable can be approximated by a $Poisson(\lambda)$ random variable for large n and small p since

$$[(1-p) + pz]^n = [1 + \frac{\lambda(z-1)}{n}]^n \approx e^{\lambda(z-1)} \tag{43}$$

as $n- > \infty$. Also this suggests the approximation

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{(np)^k e^{-np}}{k!} \tag{44}$$

for n large, p small

Example Suppose that a dog is imperfect in counting pigs and independently identifies each incoming pig with probability p. If the distribution of the number of incoming pigs in a unit time is a Poisson distribution with parameter λ. What's the distribution of the number of counted pigs?
Soln:

$$\begin{aligned}
P(X = k) &= \sum_{n=0}^{\infty} P(N = n) P(X = k | N = n) \\
&= \sum_{n=k}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lambda^k \frac{e^{-\lambda} p^k}{n!} \sum_{n=k}^{\infty} \binom{n}{k} \lambda^{n-k} (1-p)^{n-k} \\
&= (\lambda p)^k e^{-\lambda} \sum_{n=k}^{\infty} \lambda^{n-k} \frac{(1-p)^{n-k}}{(n-k)!k!} \\
&= \frac{\lambda p)^k e^{-\lambda}}{k!} \sum_{n=k}^{\infty} \lambda^{n-k} \frac{(1-p)^{n-k}}{(n-k)!} \\
&= \frac{\lambda p)^k e^{-\lambda}}{k!} \sum_{j=0}^{\infty} \frac{\lambda^j (1-p)^j}{j!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda p}
\end{aligned} \tag{45}$$

Here $P(N = n)$ follows a Poisson distribution and $P(X = k|N = n)$ follows Binomial distribution.

2.7 The Weak law of large numbers

Let $X_1, X_2, ..., X_n$ be a sequence of uncorrelated random variables with the same mean m and the same variance σ^2. Then for every $\epsilon > 0$,

$$\lim_{n->\infty} P(|M_n - m| \geq \epsilon) = 0 \tag{46}$$

s where $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. This justifies the use of sample mean as an estimate of the true mean. It can be derived using the chebyshev inequality

$$P(|M_n - m| \geq \epsilon) \leq \frac{var(M_n)}{\epsilon^2} \tag{47}$$

and

$$var(M_n) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n} \tag{48}$$

2.8 Conditional Probability

For conditional probability distributions, we use the notation

$$P(X \in B|Y \in C) = \frac{P(X \in B, Y \in C)}{P(Y \in C)} \tag{49}$$

For conditional probability mass function, we use the notation

$$p_{x_i|y_j}(y_i|x_i) = \frac{p(x_i, y_j)}{p(x_i)} \tag{50}$$

The bayes rule gives,

$$p_{Y|X}(y_i|x_i) = \frac{p_{X|Y}(x_i, y_j)p_Y(y_j)}{p_X(x_i)} \tag{51}$$

The law of total probability gives,

$$P(Y \in C) = \sum_i P(Y \in C|X = x_i)P(X = x_i) \tag{52}$$

As an example consider the binary channel receiver design with input X and output Y and channel described by the transition probabilities given as

$$P(Y = 1|X = 0) = p_{10}, P(Y = 1|X = 1) = p_{11}$$

$$P(Y = 0|X = 0) = p_{00}, P(Y = 0|X = 1) = p_{00}$$

Consider the problem of the receiver design, The receiver has access to the channel output Y and must estimate the value of X. The decision rule with the the smallest probability of error is the maximum aposteriori (MAP) rule. Having observed $Y = j$ the MAP rule says to decide $X = 1$ if

$$P(X = 1|Y = j) \geq P(X = 0|Y = j) \tag{53}$$

and decide $X = 0$ otherwise. Using the bayes rule the above relation simplifies to

$$P(Y = j|X = 1)P(X = 1) \geq P(Y = j|X = 0)P(X = 0) \tag{54}$$

Or,

$$\frac{P(Y = j|X = 1)P(X = 1)}{P(Y = j|X = 0)P(X = 1)} \geq \frac{P(X = 0)}{P(X = 1)} \tag{55}$$

The left hand side of the relation is called the likelihood ratio and the right hand side is the threshold which is 1 for the equiprobable case.

2.9 Conditional Expectation

The conditional expectation is given as

$$E[g(Y)|X = x_i] = \sum_j g(y_j)P_{Y|X}(y_i|x_i) \tag{56}$$

The substitution law for the conditional expectation is

$$E[g(X, Y)|X = x_i] = E[g(x_i, Y)|X = x_i] \tag{57}$$

The law of total probability for expectation is

$$E[g(X, Y)] = \sum_i E[g(X, Y)|X = x_i]P_X(x_i) \tag{58}$$

and if g is a function of Y only

$$E[g(Y)] = \sum_i E[g(Y)|X = x_i]P_X(x_i) \tag{59}$$

Note that if $h(x_i) = E[g(X, Y)|X = x_i]$ then

$$E[g(X)] = \sum_i h(x_i)P_X(x_i) = \sum_i E[g(X, Y)|X = x_i]P_X(x_i) = E[g(X, Y)] \tag{60}$$

i.e. h(X) is a random variable whose expectation is $E[g(X, Y)]$.

3 Continuous Random Variables

3.1 Densities and Probabilities

X is called a continuous random variable if $P(X \in B) = \int_B f(t)dt$ for some integrable function f which is non-negative and integrates to one over the real line i.e. $\int f(t)dt = 1$. A non-negative function that integrates to one is called a probability density function (pdf). If the set B is an interval such that $B = [a, b]$,

$$P(a \leq X \leq b) = \int_a^b f(t)dt \tag{61}$$

A random variable with uniform pdf is used to model experiments in which the outcome is constrained to lie in a known interval say $[a, b]$ and all outcomes are equally likely. Write f uniform[a,b] if $a < b$ and

$$f(x) = \frac{1}{b-a} \tag{62}$$

An exponential random variable is used to model lifetimes such as how long a lightbulb burns. If X is exponentially distributed with parameter λ,, we write X $exp(\lambda)$ and the pdf is

$$f(x) = \lambda e^{-\lambda x} \tag{63}$$

If U uniform (0,1) then $X = ln(1/U)$ is exp(1)

A double sided exponential is also called Laplace. It models the difference of two independent exponential random variables. If X is a laplacian random variable, we write X Laplace(λ) and the pdf is

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x|} \tag{64}$$

The Cauchy random variable models the quotient of independent gaussian random variables. If X is a cauchy random variable with parameter $\lambda > 0$, we write X Cauchy (λ) and the pdf is

$$f(x) = \frac{\lambda / pi}{\lambda^2 + x^2} \tag{65}$$

The gaussian or normal random variable is used to model the sum of many independent random variables. They arise as noise models in communication and control systems. If X is a gaussian random variable with mean μ and variance σ^2, we write X $N(\mu, \sigma^2)$ and the pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{66}$$

If $\mu = 0$ and $\sigma^2 = 1$, we say f is a standard normal density. To show that normao density integrates to one i.e. $I = \int e^{-x^2/2} dx = \sqrt{2\pi}$, the trick to show $I^2 = 2/pi$. The sum of squares of two independent normal random variables, $U^2 + V^2$ is exponential.

The rayleigh density is given as

$$f(x) = x e^{-x^2/2} \tag{67}$$

If U and V are independent normal random variables, $\sqrt{U^2 + V^2}$ is rayleigh distributed. If f is a probability density function then for a real number c and a positive number λ the function given by

$$\lambda f(\lambda(x - c)) \tag{68}$$

is also a pdf and c is called the location parameter and λ is called the scale parameter. In the exponential, laplace and cauchy densities the parameter λ is a scale parameter while in a gaussian density $c = \mu$ is the location parameter and $\lambda = 1/\sigma$ is the scale parameter.

The basic gamma density with parameter $p > 0$ is given by

$$g_p(x) = \frac{x^{p-1} e^{-x}}{\Gamma(p)} \tag{69}$$

where

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx \tag{70}$$

is the gamma function.

The general gamma density is defined as $g_{p,\lambda}(x) = \lambda g_p(\lambda_x)$ with the following special cases of great important:

- If $p = m$ is a positive integer, $g_{m,\lambda}$ is called an $Erlang(m, \lambda)$ density. The sum of m i.i.d. $\exp(\lambda)$ random variables is an $Erlang(m, \lambda)$ random variable.
- If $p = k/2$ and $\lambda = 1/2$, $g_{p,\lambda}$ is called a chi-squared density with k degrees of freedom. It arises as the square of a normal random variable.

3.2 Transform Methods

Recall that the probability generating function applies only to non-negative integer valued random variables. The moment generating function (mgf) of a real-valued random variable X is defined as

$$M_X(s) = E[e^{sX}] \tag{71}$$

and generalizes the concept of probability generating function since if X is discrete taking only non-negative integer values, then

$$M_X(s) = E[e^{sX}] = E[(e^s)^X] = G_X(e^s) \tag{72}$$

The moment generating function when differentiated k times and evaluated at $s = 0$ gives the k^th moment

$$M_X^{(k)}(s)|_{s=0} = E[X^k] \tag{73}$$

assuming $M_X(s)$ is finite in a neighborhood of $s = 0$ If X is a continuous random variable with density f, then the mgf is the laplace transform of f evaluated at $-s$.

$$M_X(s) = E[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f(x)dx \tag{74}$$

If $M_X(s)$ is finite for all real s in a neighborhood of the origin, say for $-r < s < r$ for some $0 < r < \infty$ then X has finite moments of all orders and the following calculation using the power series $e^\epsilon = \sum_{n=0}^{\infty} \epsilon^n/n!$ is valid for complex s with $|s| < r$,

$$E[e^{sX}] = E[\sum_{n=0}^{\infty} \frac{(sX)^n}{n!}] = \sum_{n=0}^{\infty} \frac{s^n}{n!} E[X^n] \tag{75}$$

The moment generating function may not be defined for all values of s for some random variables, for instance $M_X(s) = \infty$ for all real $s \neq 0$ for a cauchy random variable. A characteristic function however always exists. The characteristic function of X is defined as

$$\phi_x(\mu) = E[e^{j\mu X}] \tag{76}$$

Note that $\phi_x(\mu) = M_X(j\mu)$. Also, since $|e^{j\mu X}| = 1$, $|\phi_X(\mu)| \leq |e^{j\mu X}| = 1$. Hence the characteristic function always exists and is bounded in magnitude by one. Note that if X is a continuous random variable with density f then the characteristic function is simply the Fourier transform of the pdf.

$$\phi_X(\mu) = \int_{-\infty}^{\infty} e^{j\mu X} f(x)dx \tag{77}$$

The inverse Fourier transform of the characteristic function gives the pdf,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-j\mu X} \phi_X(\mu)d\mu \tag{78}$$

And if X is an integer valued random variable then the characteristic function is a 2π-periodic Fourier series.

$$\phi_X(\mu) = E[e^{j\mu X}] = \sum_n e^{j\mu n} P(X = n) \tag{79}$$

And given the characteristic function the Fourier series coefficients are

$$P(X = n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-j\mu n} \phi_X(\mu) d\mu \tag{80}$$

The moments can be obtained by the characteristic function as follows,

$$\phi_X^{(k)}(\mu)|_{\mu=0} = j^k E[X^k] \tag{81}$$

assuming $E[X^k] < \infty$

3.3 Expectation of multiple random variables

If X and Y are independent continuous random variables then the characteristic function of their sum $Z = X + Y$ is the product of the characteristic functions of X and Y

$$\phi_Z(\mu) = E[e^{j\mu X}]E[e^{j\mu Y}] = \phi_X(\mu)\phi_Y(\mu) \tag{82}$$

and the density is the convolution of their densities,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy \tag{83}$$

Using the fact that $X \leq |X|$ it can be shown that $|EX| \leq E[|X|]$

3.4 Probability bounds

A result known as Chernoff bound, much stronger than Markov and Chebyshev Inequalities is given by

$$P(X \geq a) \leq inf_{s>0}[e^{-sa} M_X(s)] \tag{84}$$

where the infimum is over all $s > 0$ for which $M_X(s)$ is finite. For sufficiently large a the Chernoff bound on $P(X \geq a)$ is always smaller than the bound obtained by the Chebyshev inequality which is smaller than the one obtained by Markov inequality.

4 Cumulative Distribution Functions

4.1 Continuous random variables

The cummulative distribution function of a continuous random variable X is defined as

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{\infty} f(t) dt \tag{85}$$

The cdf of a gaussian random variable cannot be expressed in a closed form. If $X\ N(\mu, \sigma^2)$ then

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dt \tag{86}$$

which can be expressed using the standard gaussian cdf

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \tag{87}$$

to get

$$F(x) = \Phi(\frac{x-\mu}{\sigma}) \tag{88}$$

Since the gaussian cdf cannot be expressed in a closed form it can be expressed in terms of the error function which is computed numerically and is given as

$$erf(y) = \frac{2}{\sqrt{\pi}} \int_{0}^{y} e^{-t^2} dt \tag{89}$$

Clearly $\Phi(y) = \frac{1}{2}[1 + erf(y/\sqrt{2})]$. And the function $Q(y) = 1 - \Phi(y)$ can be expressed in terms of the complimentary error function

$$erf(y) = 1 - erf(y) = \frac{2}{\sqrt{\pi}} \int_{y}^{\infty} e^{-t^2} dt \tag{90}$$

to get

$$Q(y) = \frac{1}{2} erfc(\frac{y}{\sqrt{2}}) \tag{91}$$

The main application of the cdf is in finding the probability density of $Y = g(X)$ given function g and the density of X. This happens when the input to a system g is modeled as a random variable and the system output is another random variable and we would like to compute probabilities involving Y. Clearly $P(X > x) = 1 - F_X(x)$ and the density can be recovered from the cdf by simple differentiation.

$$f(x) = F'(x) \tag{92}$$

If the function $g(x)$ is continuous and strictly increasing then assuming $h(y) = g^{-1}(y)$ we have

$$f_Y(y) = f_X(h(y)|h^{'}(y)) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} \tag{93}$$

The conditional cdf of Y given X is given by

$$F_{Y|X}(y|x_i) = P(Y \leq y|X = x_i) \tag{94}$$

If $F_{Y|X}(y|x_i)$ is differential with respect to y, this derivative is called conditional density of Y given X and is denoted by $f_{Y|X}(y|x_i)$.

To simulate a random variable Y with the cdf $F(y)$ generate a uniform random variable $X\quad$ uniform[0,1] and apply the transformation $Y = F^{-1}(X)$. The cdf of Y then is

$$F_Y(y) = P(Y \leq y) = P(F^{-1}(X) \leq y) = P(X \leq F(y)) = F(y) \tag{95}$$

4.2 Discrete Random Variables

For a discrete random variable taking distinct values x_i,

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} P(X = x_i) \tag{96}$$

and for two adjacent values $x_{j-1} < x_j$

$$P(X = x_j) = F(x_j) - F(x_{j-1}) \tag{97}$$

and for $x_{j-1} \leq x \leq x_j$, $F(x) = F(x_{j-1})$.

4.3 Mixed Random Variables

A random variable with density involving impulse terms in called a mixed random variable and the density is said to e impulsive or the generalized density. It is of the form

$$f_Y(y) = f_Y(y) + \sum_i P(Y = y_i)\delta(y - y_i) \tag{98}$$

where the y_is are distinct points at which $F_Y(y)$ has jump discontinuities. If X is a mixed random variable then $E[k(X)]$ is given by

$$E[k(X)] = \int_{-\infty}^{\infty} k(x)f(x)dx = \int_{-\infty}^{\infty} k(x)f_X(x) + \sum_i k(x_i)P(X = x_i) \tag{99}$$

4.4 Functions of random variables

Consider systems modeled by a real-valued function $g(x)$. The system input is a random variable X and the system output is the random variable $Y = g(x)$. $F_Y(y)$ is then given as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \in B_y) \tag{100}$$

where $B_y = \{x \in R | g(x) \leq y\}$. And if X has the density $f_X(x)$ then

$$F_Y(y) = P(X \in B_y) = \int_{B_y} f_X(x)dx \tag{101}$$

The set B_y can usually be identified by sketching the function $g(x)$.

4.5 Properties of Cdfs

Let X be a real-valued random variable with cdf $F(x)$. Then F satisfies the following:

- $0 \leq F(x) \leq 1$
- if $a < b$ then $P(a < X \leq b) = F(b) - F(a)$
- F is nondecreasing

- $lim_{x->\infty}F(x) = 1$
- $lim_{x->-\infty}F(x) = 0$
- $P(X = x_0) = F(x_0) - F(x_0-)$

Also note that $P(X > x_0) = 1 - F(x_0)$ and $P(X/gex_0) = 1 - F(x_0-)$. However if F is continuous at x_0, $F(x_0-) = F(x_0)$ and so $P(X \geq X_0) = 1 - F(x_0)$ and $P(X = x_0) = 0$.

4.6 The Central Limit Theorem

Let $X_1, X_2, ...$ be independent identically distributed random variables with finite mean μ and finite variance σ^2. If Y_n is defined as

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma}) \tag{102}$$

which has zero mean and unit variance, then

$$lim_{n->\infty}F_{Y_n}(y) = \Phi(y) \tag{103}$$

where $\Phi(y) = \int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$ is the standard normal cdf. In practice, for fixed n, the approximation is better for values of y near the origin than for values of y too far away from the origin. The derivation of the center limit theorem is as follows: Assume $X_1, X_2, ..., X_n$ is a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $Z_i = (X_i - \mu)/\sigma$. The Z_i are zero mean and unit variance i.i.d. random variables with the commom characteristic function $\phi_Z(\mu) = E[e^{j\mu Z_i}]$. Since,

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i \tag{104}$$

we can write the characteristic function of Y_n as

$$\phi_{Y_n}(\nu) = E[e^{j\nu Y_n}] = E[e^{(j\frac{\nu}{\sqrt{n}}\sum_{i=1}^{n}Z_i)}] = \phi_Z(\frac{\nu}{\sqrt{n}})^n \tag{105}$$

And using the expansion for the exponential function

$$\phi_Z(\frac{\nu}{\sqrt{n}}) = E[e^{j(\nu/\sqrt{n})}Z_i] = E[1+j\frac{\nu}{\sqrt{n}}Z_i+\frac{1}{2}(j\frac{\nu}{\sqrt{n}}Z_i)^2+R(j\frac{\nu}{\sqrt{n}}Z_i)] = 1+0-\frac{1}{2}(\frac{\nu^2}{n})^2+E[R(j\frac{\nu}{\sqrt{n}}Z_i)] = 1-\frac{\nu^2/2}{n} \tag{106}$$

Hence,

$$\phi_{Y_n}(\nu) = \phi_Z(\frac{\nu}{\sqrt{n}})^n = (1 - \frac{\nu^2/2}{n})^2 - > e - \nu^2/2$$

And since the characteristic function converges to that of the standard gaussian random variable, so does the cdf $F_{Y_n}(y) - > \Phi(y)$

4.7 Reliability

Let T be the lifetime of a device or system. The reliability function of the device or system is defined by

$$R(t) = P(T > t) = 1 - F_T(t) \tag{107}$$

which implies that the reliability at time t is the probability that the lifetime is greater than t.

The mean time to failure (MTTF) is the expected lifetime, $E[T]$. Since lifetimes are non-negative.

$$E[T] = \int_0^\infty P(T > t)dt = \int_0^\infty R(t)dt \tag{108}$$

The failure rate of a device or system with lifetime T is

$$r(t) = lim_{\delta t->0} \frac{P(T \leq t + \delta t | T > t)}{\delta T} = -\frac{R(t+\delta t) - R(t)}{R(t)\delta t} = -\frac{R'(t)}{R(t)} = \frac{f_T(t)}{\int_t^\infty f_T(\theta)d\theta} \tag{109}$$

So the failure rate is completely determined by the density f_T and the converse is also true,

$$f_T(t) = r(t)e^{-\int_0^t r(\tau)d\tau} \tag{110}$$

5 Bivariate random variables

5.1 Joint and Marginal Probabilities

The main focus here is the study of pairs of continuous random variables that are not independent. Consider the following functions of two random variables X and Y

$$X + Y, XY, max(X, Y), min(X, Y)$$

These forms appear in practice-in a telephone channel the signal X is corrupted by additive noise Y; in a wireless channel the signal X is corrupted by fading i.e. multiplicative noise; if X and Y are the traffic rates at two different routers of an ISP then we desire $max(X, Y) \leq u$ where u is the router capacity; if X and Y are sensor voltages we may need to trigger the event $min(X, Y) \leq v$ for some threshold v. THe cdfs of these functions can be expressed in the form $P(X, Y) \in A$ for various sets $A \in R^2$.

5.1.1 Product set and marginal probabilities

The cartesian product of two univariate sets B and C is defined as

$$B \times C = \{(x, y) | x \in B and y \in C\} \tag{111}$$

Thus,

$$P(X \in B, Y \in C) = P9(X, Y) \in B \times C) \tag{112}$$

The marginal probability can then be written as

$$P(X \in B) = P((X, Y) \in B \times R) \tag{113}$$

5.1.2 Joint and marginal Cumulative Distribution Functions

The joint cumulative distribution function of X and Y is defined by

$$F_{X,Y}(x, y) = P(X \leq X, Y \leq y) = P((X, Y) \in (-\infty, x] \times (-\infty, y]) \tag{114}$$

The joint cdf lets us compute $P((X,Y) \in A)$ for any set A. In particular, consider the rectangle formula. The $P(a < X \leq b, c < Y \leq d)$ is given by

$$F_{XY}(b,d) - F_{XY}(a,d) - F_{XY}(b,c) + F_{XY}(a,c) \tag{115}$$

The marginal cdf F_X can be obtained from the joint cdf as

$$F_X(x) = lim_{y->\infty} F_{XY}(x,y) = F_{XY}(x,\infty) \tag{116}$$

5.1.3 Independent Random Variables

The random variables X and Y are independent if and only if for all sets B and C,

$$P(X \in B, Y \in C) = P((X,Y) \in B \times C) = P(X \in B)P(Y \in C) \tag{117}$$

or in other words, if and only if the join cdf factors,

$$F_{XY}(x,y) = F_X(x)F_Y(y) \tag{118}$$

5.2 Jointly Continuous Random Variables

The random variables X and Y are jointly continuous with joint density $f_{XY}(x,y)$ if

6 Distribution Derived From the Normal

6.1 chi-square,t, and F distributions

6.1.1 Chi-square distribution

If $Z \sim N(0,1)$, the distribution of $U = Z^2$ is called the chi-square distribution with 1 degree of freedom, χ_1^2

For chi-square random variable with n degree of freedom, χ_n^2 distribution can be understood as the distribution that results from summing the squares of n independent random variables $\sim N(0,1)$

$$V = \sum_i^n Z_i^2$$

6.1.2 t distribution

If $U_1, U_2, ..., U_n$ are independent chi-square random variables with 1 degree of freedom, the distribution of $V = U_1 + U_2 + ... + U_n$ is called the chi-square distribution with n degrees of freedom, χ_n^2.Now, if $Z \sim N(0,1)$ and $V \sim \chi_n^2$ and Z and V are independent, then the distribution of $Z/\sqrt{V/n}$ is called the t distribution with n degree of freedom. t distribution is very close to standard normal distribution for more than 20 or 30 degrees of freedom.

6.1.3 F distribution

F distribution is based on chi-square random variables. If X and Y are independent chi-square variables with m and n degrees of freedom, respectively. The distribution of $W = \frac{X/m}{Y/n}$ is called the F distribution with m and n degrees of freedom, $F_{m,n}$

6.2 Degree of Freedom

1. Why sample variance has $n-1$ in the denominator?

This $n-1$ is the degree of freedom (DOF), representing the number of independent pieces of information on which the estimate is based. In general, the DOF for *an estimate* is equal to the number of observations/r.v.s in the sample minus the number of parameters estimated en route to the final parameter that we want to estimate. For example, we have two random variables in a sample ($X_1 = 8, X_2 = 5$) and we want to find the sample variance s^2 as an estimate of σ^2. But before we can calculate the sample variance, we had to estimate the sample mean ($\hat{\mu}$) because we need $\hat{\mu}$ to calculate the sample variance. Therefore, the estimate of the population variance has $2 - 1 = 1$ DOF. If we had 100 observations/r.v.s in our sample, then our estimate of variance would have had 99 DOF. Therefore, the DOF of an estimate of variance is equal to $n-1$, where n is the number of observations.

$$s^2 = \frac{\sum_i (X_i - \hat{\mu})^2}{n-1}$$

7 Survey Sampling

8 Parameter Estimation

8.1 Parameter Estimation

1. Why does one need to fit probability laws to data?
The form of a probability distribution and the parameters of that distribution may be of scientific interest
For descriptive purposes as a method of data summary or compression
Distributions once generated, can be used in various simulation (pricing, planning)
2. The observed data will be regarded as realizations of random variables, $X_1, X_2, , X_n$ with a joint distribution $f(x|\theta)$; an estimate of theta, will be a function of $X_1, X_2, , X_n$ and is thus a random variable with a probability distribution (sampling distribution). We often use standard error to quantify the variability of the estimate.
3. Two approaches for forming estimates
a. the method of moments
b. the method of maximum likelihood (generally useful)
4. Advanced theory of statistics is heavily concerned with optimal estimation. The essential idea is that wed like to chose the estimation procedure that gives an estimate whose sampling distribution is most concentrated around the true parameter value.

8.2 The Method of Moments

1. What's a moment of *a probability law* and its relation to the moment-generating function?
The kth moment of *a probability law* is defined as: $\mu_k = E(X^k)$, where X is a random variable following that probability law. If $X_1, X_2, , X_n$ are i.i.d. random variables (drawn multiple times) from that distribution, the kth sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_i X_i^k$$

For example, we can write $\sum_i X_i^2$ in terms of the 2nd sample moment, i.e., $\sum_i X_i^2 = n\hat{\mu}_2$. We can view $\hat{\mu}_k$ as an estimate of μ_k, which is equal to $E[X^k]$.
The moment-generating function of a random variable X is $M(t) = E(e^{tX})$ and in a discrete case, $M(t) = \sum_x e^{tx}p(x)$
And their relation is
$M^{(k)}(0) = \mu_k = E(X^k)$
2.What's the essence of the method of moments?
The MoM estimates parameters by finding expressions for them in terms of the lowest possible order moments and then substituting sample moments into the expressions.
3. Stepwise, what is the procedure of MOM?
Say if there are two parameters in the distribution that you want to fit.
The first step is to express these parameters using real moments:
$\theta_1 = f_1(\mu_1, \mu_2)$ $\theta_2 = f_2(\mu_1, \mu_2)$
The next step is to use sample moments to substitute the real moments:
$\theta_1 = f_1(\hat{\mu_1}, \hat{\mu_2})$ $\theta_2 = f_2(\hat{\mu_1}, \hat{\mu_2})$
Usually the first step involves finding expressions for the moments in terms of the parameters and then invert. 4. How stale is the estimate?
Standard statistical technique is to derive the sampling distribution of the estimate or an approximation to that distribution.
As a concrete example, say we have a number of samples X_i, which are independent Poisson random variables with parameter λ_0. We want to estimate the parameter λ. Naturally, we'd use the first moment of the Poisson distribution to do so, namely $\hat{\lambda} = E(X) = \frac{\sum X_i}{n}$. Now if we let $S = \sum X_i$, then the parameter estimate $\hat{\lambda} = \frac{S}{n}$ would be a random variable, the distribution of which is called its sampling distribution. Note that, the distribution of the sume of independent Poisson random variables is Poisson distributed, so the distribution of S is Poisson($n\lambda_0$). Thus, the sampling distribution's pdf is

$$P(\hat{\lambda}) = P(S = n\hat{\lambda}) = \frac{(n\lambda_0)^{n\hat{\lambda}}e^{-n\lambda_0}}{n\hat{\lambda}}$$

Since S is Poisson, its mean and variance are both $n\lambda_0$, so

$$E(\hat{\lambda}) = \frac{1}{n}E(S) = \lambda_0$$

$$Var(\hat{\lambda}) = \frac{1}{n^2}Var(S) = \frac{\lambda_0}{n}$$

Now it comes the prerequisite, if $n\lambda_0$ is large (how large is large so Poisson becomes normal? λ needs to be around 1000), we know the distribution of S is approximately normal and so is $\hat{\lambda}$ with mean and variance above.
5. What do you mean when you say the estimate is unbiased? It means that the sampling distribution is centered at the true parameter, in the above example, λ_0, or $E(\hat{\lambda}) = \lambda_0$
6. What is standard error in this example?

The standard deviation of the sampling distribution is called the standard error. Specifically, the standard error of $\hat{\lambda}$, which is the statistic that we are estimating.

$$\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda_0}{n}}$$

7. What is the estimated standard error? We normally do not know the sampling distribution or the standard error because we do not know λ_0, the true value of the parameter that we want to estimate. So we use the sample estimate, $\hat{\lambda}$, to assess the variability of our estimate.

$$s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}$$

8.3 The Method of Maximum Likelihood

1. What's the essence of the method of maximum likelihood?

We want to find a parameter θ that maximizes $P(data|\theta)$ by taking derivative w.r.t. it and setting the derivative to 0; Suppose that random variables $X_1, ..., X_n$ have a joint density function $f(x_1, x_2, ..., x_n)$. Given observed values $X_i = x_i$, where $i = 1...n$, the likelihood of θ as a function of $x_1, ..., x_n$ is

$$lik(\theta) = f(x_1, x_,2, ..., x_n|\theta)$$

8.3.1 Large Sample Theory for Maximum Likelihood Estimates

Large sample theory states that using a large sample size, the sampling distribution of the maximum likelihood estimate is approximately normal with mean θ_0 and variance $1/[nI(\theta_0)]$. However, this is just a limiting result, which means that it holds as the sample size tends to infinity, and we say that the mle is asymptotically unbiased and refer to the variance of the limiting normal distribution as the asymptotic variance of the mle, where $I(\theta_0)$ is specified as follows:

$$I(\theta_0) = E[\frac{\partial}{\partial\theta_0} logf(X|\theta_0)]^2 = -E[\frac{\partial^2}{\partial\theta_0^2} logf(X|\theta_0)]$$

This is also called the **Fisher Information**. Note that here the X is just one random variable not a random vector. In other words, $f(X|\theta_0)$ is just the pdf of the underlying random variable not the likelihood function. The Cramer-Rao lower bound on the variance of the estimator is just $1/nI(\theta_0)$. But if we substitute $f(X|\theta_0)$ by the likelihood of the data,$f(\vec{X}|\theta_0)$. The Cramer-Rao lower bound on the variance of the estimator becomes $1/I(\theta_0)$. An example problem setup is as follows:

For an i.i.d. sample of size n, the log likelihood is

$$l(\theta) = \sum_i logf(x_i|\theta)$$

We denote the true value of θ by θ_0. And, the large sample theory says that under reasonable conditions $\hat{\theta}_{mle}$ is a consistent estimator of θ_0; that is, $\hat{\theta}_{mle}$ converges to θ_0 in probability as n approaches infinity.

8.3.2 Confidence Intervals for Maximum Likelihood Estimates

A confidence interval for θ is an interval based on the sample values used to estimate θ. That is, we develop confidence interval for θ based on $\hat{\theta}$. Since these sample values are random, the interval is thus also random and the probability that it contains θ is called the coverage probability of the interval. For example, a 95 percent confidence interval for θ is a random interval that contains θ with probability 0.95. A confidence interval quantifies the uncertainty of a parameter estimate.

There are three methods to form confidence intervals for mles.

- Exact methods
- Approximations based on the large sample properties of mles
- Bootstrap confidence intervals

Exact method

For an normally i.i.d. sample $X_1, X_2, ..., X_n$, we know the maximum likelihood estimates of μ and σ are as follows:

$$\hat{\mu} = \bar{X}$$

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

The exact method assumes that the confidence interval for μ is based on the fact that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

Where S is the population standard deviation, i.e., $S^2 = \sigma^2$. So based on this, we can have

$$P(-t_{n-1}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{n-1}(\alpha/2)) = 1 - \alpha$$

From this, we can manipulate the formula to get the confidence interval as follows:

$$P(\bar{X} - \frac{S}{\sqrt{n}}t_{n-1}(\alpha/2) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}}t_{n-1}(\alpha/2)) = 1 - \alpha$$

According to the equation, the probability that μ lies in the interval is $1 - \alpha$. Note that the confidence interval is **random**.

Now let's calculate the confidence interval for σ^2. We know that

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Again, we have

$$P(\chi^2_{n-1}(1 - \alpha/2) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-1}(\alpha/2)) = 1 - \alpha$$

And after manipulation we can find the confidence interval for σ^2.

But note that exact methods such as the one illustrated here are the exception rather than the rule in practice for constructing confidence intervals. To construct an exact confidence interval we need detailed knowledge of the sampling distribution. In other words, we need to express the estimated quantify in terms of some random variable of which we know its distribution law so we can express its variance. For example, $\hat{\theta} = X/n$. The variance of $\hat{\theta}$, which is what we usually are interested, is equal to the variance of X divided by n^2. But you need to know closed form distribution of X so you know its variance. X is better be of some familiar distribution like binomial, Poisson, etc. A second method of constructing confidence interval is based on the large sample theory.

Approximations based on the large sample properties of mles

If we can't know the sampling distribution like we did for X above, then the large sample properties of mles comes into play. According to the large sample theory, the distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ is approximately the standard normal distribution. Since θ_0 is unknown, we will use $I(\hat{\theta})$ in place of $I(\theta_0)$. Again, apply the confidence interval formula, we get

$$P(-z(\alpha/2) \leq \sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \leq z(\alpha/2)) \approx 1 - \alpha$$

where $z(.)$ is critical value for the standard normal distribution, i.e., 1.96 for 95 percent confidence interval.

Bootstrap confidence intervals

Again, suppose that $\hat{\theta}$ is an estimate of the true parameter θ_0 and suppose that we know the distribution of their difference $\Delta = \hat{\theta} - \theta_0$. So we can express the CI as follows:

$$P(\delta_{lowerPercentile} \leq \hat{\theta} - \theta_0 \leq \delta_{upperPercentile}) = 1 - \alpha$$

From this we can then calculate the confidence interval for θ_0. But this assumes that the distribution of $\Delta = \hat{\theta} - \theta_0$ is known, which is typically not the case because θ_0 is not known. But as before we could use $\hat{\theta}$ in place of θ_0 and generate many samples from a distribution with value $\hat{\theta}$. And, from each sample construct an estimate of θ, say θ^*. Then the distribution of $\hat{\theta} - \theta_0$ is approximated by that of $\theta^* - \hat{\theta}$. And the quantiles of which are then used to form an approximate confidence interval.

8.3.3 The Bayesian Approach to Parameter Estimation

1. What's the Bayesian approach to parameter estimation and how it's different from the Frequentist approach?

Suppose that you have a sample of random variables $X_1, X_2, ..., X_n$, and you are trying to estimate a parameter, θ, of the distribution from which the sample was drawn (each item of the sample was individually drawn if the sample is said to be i.i.d.). In Bayesian approach, you would assume that the unknown parameter θ is treated as *a random variable* which has a distribution, called the prior distribution, representing what you know about the parameter before observing the sample. This framework is in contrast with the approaches (often called the Frequentist approach) described in previous sections, in which θ was treated as an unknown *constant*. For a given value, $\Theta = \theta$, the data now have the probability distribution $f_{X|\Theta}(x|\theta)$. And, the joint probability of X and Θ is thus,

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x|\theta) f_\Theta(\theta)$$

2. What's the poster distribution and what does it represent for?

The posterior distribution is:

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} = \frac{f_{X,\Theta}(x, \theta)}{\int_\theta f_{X,\Theta}(x, \theta) d\theta}$$

It represents what we now know about Θ after we observer the data X.

3. What's the posterior mode?

The most probable value of Θ in $f_{\Theta|X}(\theta|x)$.

4. Given a variety of possible estimates, how would you choose which to use?

Chose one whose sampling distribution is most highly concentrated about the true parameter value by mean squared error.

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2] = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta_0)^2$$

If the estimate $\hat{\theta}$ is unbiased which means $E[\hat{\theta}] = \theta_0$, then the MSE is just the variance.

5. What is Cramer-Rao Inequality?

Let $X_1, X_2, ..., X_n$ be i.i.d. with density function $f(x|\theta)$. Let $T = t(X_1, .., X_n)$ be an unbiased estimator,θ,. Then, under smoothness assumptions on the density function,

$$Var(T) \geq \frac{1}{nI(\theta)}$$

6. What is a sufficient statistic and give an example?

A function $T(X_1, X_2, ..., X_n)$ is called a *sufficient statistic of* θ if all the information in the sample about θ is contained in T.

9 Testing Hypotheses and Assessing Goodness of Fit

1. What is Hypotheses testing?

Use likelihood ratio or posterior probability ratio to assess the evidence for each hypothesis.

2. What is type I error, significant level, type II error, power of the test, and rejection region?

- Type I error: reject H_0 when H_0 is true.
- Significance level: The probability of type I error.
- Type II error: accept H_0 when H_1 is true.
- Power of the test: 1-P(Type II error)
- Rejection region: The test statistics that lead to rejection of the null hypothesis are called rejection region.

3. What is the Neyman-Pearson Lemma?

All other tests (those that are not based on likelihood ratio) that can reach smaller significant levels have power less or equal to that of the likelihood ratio test. Alternatively, among all tests with significance level α, the test that rejects for small values of the likelihood ratio is most powerful.

4. Given a significance level α, how to find its corresponding rejection region?

Let $X_1, ..., X_n$ be a random sample from a normal distribution with variance σ^2. Consider two simple hypotheses: $H_0 : \mu = \mu_0$, $H_1 : \mu = \mu_1$, where μ_0 and μ_1 are given constants. If we want to find the threshold value to reject small likelihood ratios, we first need to calculate the likelihood ratio and analyze when it's small. For example,

$$\frac{f_0(\vec{X})}{f_1(\vec{X})} = \frac{exp[\frac{-1}{2\sigma^2}\sum_i (X_i - \mu_0)^2]}{exp[\frac{-1}{2\sigma^2}\sum_i (X_i - \mu_1)^2]}$$

After we cancel the multipliers of the exponential, small values of the likelihood ratio is

$$2n\bar{X}(\mu_0 - \mu_1) + n\mu_1^2 - n\mu_0^2$$

Now, if (μ_0 and μ_1 are given) $(\mu_0 - \mu_1) < 0$, the likelihood ratio is small if $\bar{X}$ is large. And the significance level is the probability of rejecting H_0 when H_0 is true. Now, let's rethink our goal which is reject H_0 for small likelihood values, which can be achieved by having a small threshold because we reject when LR is smaller than something such that $LR < c$. But since the threshold of the LR and the threshold(denote it by x_0) of $\bar{X}$ have an inverse relationship,if we want to reject H_0 for small likelihood ratios, we would then need $\bar{X} > x_0$ for a x_0 as big as possible. Because a larger x_0 for $\bar{X}$ means a smaller threshold on the actual likelihood ratio (rejection corresponds to $LR < c$). Thus, a smaller threshold would correspond to smaller likelihood ratio values. Thus, we choose x_0 to give the test the desired significance level α.

Under H_0 in this example, the null distribution of the test statistic, $\bar{X}$, is a normal distribution with mean μ_0 and variance $\frac{\sigma^2}{n}$, so x_0 can be chosen from tables of the standard normal distribution.

$$P(\bar{X} > x_0) = P(\frac{\bar{X} - \mu_0}{\sigma\sqrt{n}} > \frac{x_0 - \mu_0}{\sigma\sqrt{n}})$$

such that

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

And solve for x_0.

5. Why the Neyman-Pearson Lemma is of little direct use?

Because the case of testing a simple null hypothesis versus a simple alternative is rare in reality. More often, the hypothesis does not completely specify the probability distribution (hypothesis is vague in a sense), the hypothesis is called a composite hypothesis.

6. Do you need to know distribution of both null and alternative hypothesis to use the Neyman-Pearson Lemma?

No. One of the strengths of the NP approach is that only the distribution under the null hypothesis in needed in order to construct a test.

7. How to choose null hypotheses?

The general rule of choosing null hypothesis is that the null is always the normal one compared to the alternative and has a simpler form than the alternative. For example, we can assume that the null follows a specific form of distribution like Poisson and the alternative does not. And, the null has no extraordinary capability and the alternative dose.

8. What is uniformly most powerful tests?

As an example, a most powerful test rejects for $\bar{X} > x_0$, where x_0 only depends on μ_0, σ, n but not on the alternative hypothesis μ_1. Because the test is most powerful and is the same for every alternative, it is uniformly most powerful.

9. What's the duality of confidence interval ad hypothesis tests?

For example, we have a test $H_0 : \mu = \mu_0$ and are given a random sample $X_1, ..., X_n$ to conduct the hypothesis test. In a confidence interval sense, we can assume that μ_0 in the example is the true population mean that we want to estimate using the sample data. And, in the context of hypothesis testing, we can say that a $100(1-\alpha)$ percent confidence interval for μ_0 consists of all those values of μ_0 for which the hypothesis that $\mu = \mu_0$ will be accepted at level α. To derive such confidence interval, we first need to find the acceptance region of the test.

9.1 Specification of the Significance Level and the Concept of a p-value

The theory requires that we specify the significant level, α, in advance of analyzing the data. In practice, we are often costumed to choose small values for the probability of type I error such as 0.01 and 0.05.

For example, a man is asked to identify the suits of 20 cards drawn randomly with replacement from a 52 card deck. The rejection region for a significance level $\alpha = 0.05$ is $T \geq 8$, i.e., $P(T \geq 8) = 0.05$, meaning that the probability that we falsely reject the null hypothesis (the man has no extraordinary ability) when more than 7 cards are identified correctly is 0.05. Now, the guy actually identified 9 cards correctly (called the evidence), if now we use 9 as the threshold to calculate the significant level (probability of type I error : reject null when null is true), we would have $P(T \geq 9) = 0.041$, and we call this probability the *p-value*, which is used to summarize the evidence against the null hypothesis. I think that the p-value is to quantify the credibility of the observed data. In other words, if placing the observed data in the null hypothesis, the probability that we would falsely do something (reject H_0) is greater than 0.05, then we would question the credibility of the observed data. Normally, if the p-value is less than 0.05 we consider the observation to be trustful. In the example above, if 9 cards were identified correctly, the p=value is 0.041. And if 10 cards were identified correctly, the p-value would be 0.014, representing a lower chance of falsely rejecting the null hypothesis. That is, the smaller the p-value, the stronger the evidence against the null hypothesis.

The p-value is the probability of getting a result (think of repeating the experiment) that is equal or more extreme (i.e., get 11 cards correct) than the one that's actually observed under the condition that the null hypothesis is true.

9.2 Generalized Likelihood Ratio Tests

The generalized likelihood ratio tests are mainly for situations in which the hypotheses are not simple. Specifically, we assumes that the observed data $\vec{X} = (X_1, X_2, ..., X_n)$ has a joint density or frequency function $f(\vec{X}|\theta)$. Then H_0 specifies that $\theta \in \omega_0$, where ω_0 is a subset of the set of all possible values of θ, and H_1 specifies that $\theta \in \omega_1$, a different subset of possible values of θ. Note that ω_0 and ω_1 are disjoint. If the hypotheses are composite (a hypothesis does not completely specify the probability distribution), each likelihood is evaluated at the value of θ that maximizes it, yielding the generalized likelihood ratio

$$\Lambda' = \frac{max_{\theta \in \omega_0}[lik(\theta)]}{max_{\theta \in \omega_1}[lik(\theta)]} = \frac{max_{\theta \in \omega_0} P(\vec{X}|\theta)}{max_{\theta \in \omega_1} P(\vec{X}|\theta)}$$

Or,

$$\Lambda = \frac{max_{\theta \in \omega_0}[lik(\theta)]}{max_{\theta \in \Omega}[lik(\theta)]}$$

Where $\Omega = \omega_0 \cup \omega_1$. Here small values of Λ favors rejection of H_0, for example, $\Lambda \leq \lambda_0$. And, the threshold is chosen so that $P(\Lambda \leq \lambda_0 | H_0) = \alpha$, which is the desired significance level of the test.

1. Construct a generalized likelihood ratio test for testing a normal mean. For example, we have a sample $X_1, X_2, ..., X_n$ with pdf $N(\mu, \sigma^2)$, where σ is known. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Here μ_0 is a pre-specified scalar.

We realize that in this example, the role of θ is played by μ, and $\omega_0 = \{\mu_0\}$, $\omega_1 = \{\mu | \mu \neq \mu_0\}$, and $\Omega = \{-\infty \leq \mu \leq \infty\}$. We first calculate the numerator of Λ as ω_0 only has one member. We have

$$\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_i (X_i - \mu_0)^2}$$

For the denominator, we have a lot of μs to test and we want evaluate the likelihood using the one that maximize the likelihood. We know from previous knowledge that the μ that maximizes the likelihood is the mle of $\bar{X}$. So we have

$$\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_i (X_i - \bar{X})^2}$$

After we stack the likelihood ratio, we would get

$$\Lambda = exp(-\frac{1}{2\sigma^2}[\sum_i (X_i - \mu_0)^2 - \sum_i (X_i - \bar{X})^2])$$

Recall that we want to reject small values of Λ, that is we reject $\Lambda \leq \lambda_0$. Realize that this is equivalent for rejecting large values of

$$-2log\Lambda = -\frac{1}{\sigma^2}[\sum_i (X_i - \mu_0)^2 - \sum_i (X_i - \bar{X})^2]$$

Plug in the identity,

$$\sum_i (X_i - \mu_0)^2 = \sum_i (X_1 - \bar{X})^2 + n(\bar{X} - \mu_0)^2$$

into the formula above, we see that the likelihood ratio test rejects for large values of $-2log\Lambda = \frac{(\bar{X}-\mu_0)^2}{\sigma^2/n}$, which follows chi-square distribution with 1 degree of freedom. Also, realize that under H_0, $\bar{X} \sim N(\mu_0, sigma^2/n)$, which implies that $\sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0, 1)$ and hence its square, $\frac{(\bar{X}-\mu_0)^2}{\sigma^2/n} = -2log\Lambda \sim \chi_1^2$. Now, we know the null distribution of the test statistic $(-2log\Lambda)$, we can know find the rejection region for any significance level α. The test rejects when

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

Or,

$$|\bar{X} - \mu_0| \geq \frac{\sigma}{\sqrt{n}} z(\alpha/2)$$

As a chi-square random variable with 1 DOF is the square of a standard normal random variable.

9.3 Goodness of fit

The goal of a goodness of fit metric is to measure how precise is the estimate (using the current data/sample). In other words, do we have faith in the accuracy of the first, second, third, or fourth decimal place? To answer this questions, we could find out the variance of the estimate by using its sampling distribution (often not available). We may use bootstrap to find the approximated sampling distribution. Another way is to use chi-square. Below is an example that we can use chi-square to measure the goodness of fit but instead we will use the likelihood ratio and at the end prove that the two test statistics are approximately equal.

For example, we have a theory that states that the genotypes AA, Aa, and aa occur in a population with frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2. And we have a set of data that says in a total of 1029 people, there are 342 AA, 500 Aa, and 187 aa. Here we can assume a 3 variable multinomial distribution with $p_1 = (1-\theta)^2$,$p_2 = 2\theta(1-\theta)$,and $p_3 = \theta^2$. We can write down the likelihood of this data and calculate $\hat{\theta}_{mle}$. Note that we are assuming that our p's depend on specific forms of θ. But they really don't have to be, right? So we can design a generalized likelihood ratio test as follows:

$$\Lambda = \frac{max_{p\in\omega_0}(\frac{n!}{x_1!...x_m!})p_1(\theta)^{x_1}...p_m(\theta)^{x_m}}{max_{p\in\Omega}(\frac{n!}{x_1!...x_m!})p_1^{x_1}...p_m^{x_m}}$$

such that the denominator's p is not restricted to follow any form of theta. In other words, in H_0, $p(\theta)$ and in H_1, p is free-they can be any value as long as they sum up to 1. The x_i's are the observed counts in the m cells.

By the definition of the maximum likelihood estimate, this numerator likelihood function is maximized when $\hat{\theta}$ is the maximum likelihood estimate of θ. The corresponding probability will be denoted by $p_i(\hat{\theta})$. And, since the probabilities are unrestricted under H_1, the denominator is maximized by the unrestricted (no specific form on p to follow; it purely goes with the data) mle's, or

$$\hat{p_i} = \frac{x_i}{n}$$

The likelihood ratio is, therefore,

$$\Lambda = \frac{(\frac{n!}{x_1!...x_m!})p_1(\hat{\theta})^{x_1}...p_m(\hat{\theta})^{x_m}}{(\frac{n!}{x_1!...x_m!})\hat{p_1}^{x_1}...\hat{p_m}^{x_m}}$$

And,

$$-2log\Lambda = 2\sum_i O_i log(\frac{O_i}{E_i})$$

where $O_i = n\hat{p_i}$ and $E_i = np_i(\hat{\theta})$ denote the observed and expected counts, respectively.

Now if we use Taylor series, we get

$$-2log\Lambda \sim 2n\sum_i[\hat{p_i} - p_i(\hat{\theta})] + n\sum_i\frac{[\hat{p_i} - p_i(\hat{\theta})]^2}{p_i(\hat{\theta})}$$

The first term is zero since probabilities sum to 1 and the second term is just

$$\sum_i\frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})} = \sum_i\frac{(O_i - E_i)^2}{E_i} = \chi^2$$

Now we have showed that the two test statistics are approximately the same.

If we want to use Pearson's chi-square test, and therefore, χ^2 as our test statistic. We first need to know under H_0 what distribution that it follows. The null distribution of χ^2 is approximately chi-square with 1 DOF (there are two independent cells, and one parameter, θ has to be estimated on the way before estimating the final estimate, p). With a significance level of 0.05, we can then find out the rejection region for this test is $\chi^2 > 3.84$. And from the data, we have

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.0319$$

Where the E is the expected counts computed using the estimated parameter and O is the observed counts. So the null hypothesis is not rejected. And in comparison, the likelihood ratio test statistic is

$$-2log\Lambda = 0.0319$$

The two tests lead to the same conclusion.

10 Comparing Two Samples

In many experiments, the two samples may be regarded as being independent of each other. For example, a sample of subjects may be assigned to a particular treatment (using drug A), and another independent sample may be assigned to a placebo treatment. We can have several ways to construct the groups. 1) we can randomly assign individuals to the placebo and treatment groups. 2) We can create some pairing such as each person receiving the treatment were paired with an individual of similar weight in the control group.

1. What statistical model do we need to model the observations from the placebo and treatment group?

The observations from the control group are modeled as independent random variables with a common distribution, F, and the observations from the treatment group are modeled as being independpent of each other and of the controls and as having their own common distribution function, G. In many experiments, the primary effect of the treatment is to change the overall level of the responses, so that analysis focuses on the difference of means or other location parameters of F and G.

2. What distributions should you assume for F and G and how to measure the effectiveness of the drug?

For the treatment group, we can assume that a sample $X_1, ..., X_n$, is drawn from a normal distribution that has mean μ_X and variance σ^2, and that an independent sample for the placebo group, $Y_1, ..., Y_m$, is drawn from another normal distribution that has mean μ_Y and the same variance, σ^2. The effectiveness of the treatment is characterized by the difference $\mu_X - \mu_Y$. A nature estimate of $mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$, which is the mle.

3. What distribution does $\bar{X} - \bar{Y}$ follow?

Since $\bar{X} - \bar{Y}$ may be expressed as a linear combination of independent normally distributed random variables, it is normally distributed, $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, (\frac{\sigma^2}{n} + \frac{\sigma^2}{m}))$

. 4. What is the confidence interval for $\mu_X - \mu_Y$?

$$(\bar{X} - \bar{Y}) + -z(\alpha/2)\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}})$$

I.e., the true difference, $\mu_X - \mu_Y$, is 95 percent within the interval.

5. Is σ^2 generally known?

No. We use the pooled sample variance, calculated from the samples, to estimate σ^2.

$$s_P^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{(m+n-2)}$$

where $s_X^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$.

6. What is the test statistics that we should use for hypothesis testing and for forming confidence interval?

σ^2

11 Introduction to random process

A random process is a collection of random variables that are defined on a common probability space. A random process is mathematically represented by the collection

$$\{X_t, t \in I\}$$

where X_t denotes the t^{th} random variable in the process, and the index t runs over an index set I which is arbitrary. For example, consider the experiment consisting of three coin tosses. A suitable sample space would be

$$\Omega = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$$

On this sample space, we can define several random variables. For n = 1, 2, 3, put $X_n(\omega) = 0$, if the nth component of ω is T, 1, if the nth component of ω is H

Then, for example, $X_1(THH) = 0$, $X_2(THH) = 1$, and $X_3(THH) = 1$. For a fixed ω, we can plot the sequence $X_1(\omega)$, $X_2(\omega)$, $X_3(\omega)$, called the sample path corresponding to ω. Every time the sample point ω changes, the sample path also changes.

In general, a random process X_n may be defined over any range of integers n, which we usually think of as discrete time. It is also useful to allow continuous-time random processes X_t where t can be any real number, not just an integer. Thus, for each t, $X_t(\omega)$ is a random variable, or function of ω. However, as in the discrete-time case, we can fix ω and allow the time parameter t to vary. So there are two ways to think about a random process. The first way is to think of X_n or X_t as a family of random variables, which by definition, is just a family of functions of ω. The second way is to think of X_n or X_t as a random waveform in discrete or continuous time, respectively.

asdasdas

asdasdasd

From the theoretical relationship, we have

$$AUC = \frac{1}{2} + \frac{1}{2}erf(\frac{SNR}{2})$$

Rearranging the formula to express SNR in terms of a function of AUC

$$SNR = 2erf^{-1}(2AUC - 1)$$

From the fitted function, we can have

$$SNR^2 = \frac{AA * K_1}{AA * K_2 + K_3}$$

Rearranging the above function, we have

$$AA = \frac{SNR^2 * K_3}{K_1 - SNR^2 * K_2}$$

Plug in the SNR function, we finally get

$$AA = \frac{(2erf^{-1}(2AUC - 1))^2 * K_3}{K_1 - (2erf^{-1}(2AUC - 1))^2 * K_2}$$