STAT 630: Overview of Mathematical Statistics

Nikko Cleri

General Introduction to Probability and Statistics

Some basic preliminary definitions:

- **Definition**: The *population* is a collection of numbers about which one wants to draw a conclusion or make an *inference*.
- **Definition**: The *sample* is a subset of the population.
- **Definition**: *Deduction* - reasoning from general to specific.
- **Definition**: *Experiment* - some process having a number of possible outcomes, all of which are known, but whose ultimate result is not known.
- **Definition**: *Sample Space* - the set of all possible outcomes of the experiment. The sample space is denoted S and a single element of S is s.
- **Definition**: *Event* - a subset of S

Some comments about sets:

- $A \subset B$ means that the event A is a subset of B.
- The *empty set* is the set with no elements, denoted $\emptyset$
- *Countably infinite* sets are those whose elements can be put into a 1 to 1 correspondence with the integers
- A set is *uncountably infinite* if it is neither finite nor countable.

Basic definitions of set operations:

- **Definition**: *Union* - The union $A \cup B$ is an event containing all the elements that are in A not B, B not A, or both A and B. 'A or B'
- **Definition**: *Intersection* - The intersection $A \cap B$ is an event containing all elements in both A and B. 'A and B'
- **Definition**: *Complement* - The complement of event A, A^c, is the set of all elements in the sample space that are not in A. 'not A'
- **Definition**: *Mutually Exclusive* - A and B are mutually exclusive if $A \cap B = \emptyset$, meaning they have no elements in common. Several sets $A_1, ..., A_k$ are mutually exclusive if A_i and A_j are disjoint for all $i \neq j$.

These definitions are used in the formation of *DeMorgan's Laws*:

$$A_1^c \cap \cdots \cap A_k^c = (A_1 \cup \cdots \cup A_k)^c$$
$$A_1^c \cup \cdots \cup A_k^c = (A_1 \cap \cdots \cap A_k)^c$$

Definition: $P(A)$ denotes 'the probability of event A'.

The probability measure on the sample space is the function P from subsets of S to the real line satisfying the following axioms:

- $0 \leq P(A)\ \forall A$
- $P(S) = 1$
- For any infinite sequence of disjoint events A_1, A_2...:

$$P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$$

Some simple results from these axioms:

- $P(\emptyset) = 0$
- For any finite set of disjoint events $A_1, ...A_k$:

$$P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{k} P(A_i)$$

- For any event A, $P(A) = 1 - P(A^c)$. Proof:

$$\begin{aligned} P(A) + P(A^c) &= P(S) \\ &= 1 \\ P(A) &= 1 - P(A^c) \end{aligned}$$

- $P(A) \leq 1\ \forall A$. Proof:

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ P(A^c) \geq 0 &\implies 1 - P(A^c) \leq 1 \\ P(A) &\leq 1\ \forall A \end{aligned}$$

- If $A \subset B$, then $P(A) \leq P(B)$. Also note that $A \subset B$ implies that A occurring means B must also occur.
- $A = B \implies A \subset B$ and $B \subset A$.

- For any two events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof:

$$\begin{aligned} A \cup B &= A \cup (B \cap A^c) \\ \implies P(A \cup B) &= P(A) + P(B \cap A^c) \\ P(A \cup (B \cap A^c)) &= P(A) + P(B \cap A^c) \end{aligned}$$

We can also say that

$$\begin{aligned} B &= (A \cap B) \cup (A^c \cap B) \\ P(B) &= P(A \cap B) + P(A^c \cap B) \\ P(A^c \cap B) &= P(B) - P(A \cup B) \end{aligned}$$

Finally:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B \cap A^c) \\ P(A^c \cap B) &= P(B) - P(A \cup B) \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Counting Rules, Permutations, Combinations

As we discussed in the definitions at the beginning of the previous section, $\{s_i\}$ are the outcomes of an n-member sample space S. We have:

$$\begin{aligned} \sum_{i=1}^{n} P(\{s_i\}) &= 1 \\ \because \sum_{i=1}^{n} P(\{s_i\}) &= P(S) \end{aligned}$$

Suppose we have an event $A = \{s_{i_1}, ..., s_{i_k}\}$. This implies that the probability of an event A is the sum of the probabilities of the contained outcomes:

$$P(A) = \sum_{j=1}^{k} P(\{s_{i_j}\})$$

Example: What is the probability of rolling a total of 7 on two dice?

We can count the number of outcomes that make the event of rolling a 7:

$$\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$$

And with 36 possible outcomes, the probability of rolling a 7 is $P(7) = 6/36$.

Some notes on counting rules:

Multiplication Rule: If there are r possible results of the first stage of an experiment and s possible results of the second stage and the first and second stages are independent, the total number of possible outcomes after stage 2 is $r \times s$.

A *permutation* is one possible arrangement of distinct objects. The number of permutations of n distinct objects is $n!$.

If we have a population of N distinct objects, we can draw a sample from the population in two ways:

Sampling with replacement: If we choose n objects from the population such that the object is replaced before the next object is drawn, the number of possible samples is N^n (multiplication rule n times).

Sampling without replacement: If we choose $n < N$ objects from the population such that the object is **not** replaced before the next object is drawn, the number of possible samples is

$$\begin{aligned} P_{n,N} &= N(N-1)(N-2)...(N-n+1) \\ &= \frac{N!}{(N-n)!} \end{aligned}$$

Example: What is the probability that at least two people in a group of size k have the same birthday?

The probability that no two share the same birthday would be

$$\frac{P_{k,365}}{365^k}$$

So the probability that at least 2 share the same birthday is:

$$P(k) = 1 - \frac{P_{k,365}}{365^k}$$

There is a >50% chance that two people in a room of 23 share the same birthday.

For a sample of a population of size N where the order of the elements is irrelevant, The number of unordered samples of size n is The number of ordered samples divided by the number of permutations of those samples:

$$\begin{aligned} \binom{N}{n} &= \frac{1}{n!}\frac{N!}{(N-n)!} \\ &= \frac{N!}{n!(N-n)!} \end{aligned}$$

Where the quantity $\binom{N}{n}$ is known as the *binomial coefficient*. For any real numbers x and y, the *binomial theorem* states:

$$(x+y)^N = \sum_{n=0}^{N} \binom{N}{n} x^n y^{N-n}$$

where $0! \equiv 1$

Note:

$$\begin{aligned}
\binom{N}{n} &= \binom{N}{N-n} \\
\binom{N}{n} &= \frac{N!}{n!(N-n)!} \\
\binom{N}{N-n} &= \frac{N!}{(N-n)!(N-(N-n))!} \\
&= \frac{N!}{n!(N-n)!} \\
\implies \binom{N}{n} &= \binom{N}{N-n}
\end{aligned}$$

An extension to r classes: The number of ways that n objects can be grouped into r classes with n_i in the i^{th} class is

$$\binom{n}{n_1, ..., n_r} = \frac{n!}{n_1!...n_r!}$$

Conditional Probability, Bayes' Theorem, Independence

We use the notation $P(A|B)$ to indicate the conditional probability of event A *given* event B. We define the conditional probability as

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

where it is assumed that $P(B) > 0$. The axioms of probability remain satisfied with conditional statements:

- $P(A|B) \geq) \ \forall A$
- $P(S|B) = 1$
- For every infinite sequence of disjoint events $A_1, A_2...$,

$$P(A_1 \cup A_2 \cup ...|B) = \sum_{i=1}^{\infty} P(A_i|B)$$

The multiplication rule as follows gives an alternative way of determining the intersection of two events:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

Generalized multiplication:

$$P(A_0 \cap A_1 \cap ... \cap A_n) = P(A_0 A_1 ... A_n)$$
$$P(A_0 A_1 ... A_n) = P(A_0)P(A_0|A_1) \times P(A_2|A_1 A_0) \times ... \times P(A_n|A_0 A_1 ... A_{n-1})$$

We can use this to simplify problems like the birthday problem. If we have a group of n people, what is the probability that at least two have the same birthday?

If A is the event of interest, $A^c = A_1 \cap A_2 \cap ... \cap A_n$.

$$\begin{aligned} P(A^c) &= P(A_1 ... A_n) \\ &= P(A_1)P(A_2|A_1)...P(A_n|A_1 ... A_{n-1}) \\ &= 1\left(\frac{364}{365}\right)\left(\frac{364}{363}\right)... \end{aligned}$$

Which is what we had before. The probability of at least two sharing a birthday is 1 minus this probability, which is the same as the $\frac{P_{k,365}}{365^k}$ we had from counting rules.

Example: Diagnostic testing. Define two events: D the event that the disease is present, T^+ the event that the test is positive. D^c is the event that the disease is not present and $T^- = (T^+)^c$ is the event that the test is negative. Some quantities we may care about:

- Prevalence of the disease $P(D)$
- Sensitivity of the test $P(T^+|D)$
- Specificity of the test $P(T^-|D^c)$

If we have events $B_1, B_2...$ which are mutually exclusive and exhaustive $\bigcup_{i=1}^{\infty} B_i = S$, we call $\{B_i\}$ a *partition* of the sample space S.

Definition: *The Law of Total Probability* For any event A:

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)$$

Proof: $A = A \cap S$ implies

$$\begin{aligned} A &= A \cap \left(\bigcup_{i=1}^{\infty} B_i\right) \\ &= \bigcup_{i=1}^{\infty} A \cap B_i \end{aligned}$$

Since each $A \cap B_i$ are mutually exclusive

$$\begin{aligned} P(A) &= \sum_{i=1}^{\infty} P(A \cap B_i) \\ &= \sum_{i=1}^{\infty} P(A|B_i)P(B_i) \end{aligned}$$

Definition: *Bayes' Theorem* provides a method of reversing the order of conditioning in conditional probabilities. Suppose that A and B are events with $P(A) > 0$ and $P(B) > 0$. Then we have

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

We can extend this to a situation where we know the conditional probabilities of B given each member A_i of a partition of S. Suppose $A_1, A_2...$ are mutually exclusive and exhaustive, and that $P(A_i) > 0$ for each i. Then for any event B with probability $P(B) > 0$ and for each k

$$\begin{aligned} P(A_k|B) &= \frac{P(A_k \cap B)}{P(B)} \\ &= \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)} \end{aligned}$$

Definition: We say that two events A and B are *independent* if

$$P(A \cap B) = P(A)P(B)$$

Supposing $P(B) > 0$, this is equivalent to $P(A|B) = P(A)$. Likewise for if $P(A) > 0$, we have $P(B|A) = P(B)$.

The events A and B are independent if and only if a conditional probability equals its corresponding unconditional probability.

If we have events $A_1...., A_k$, these events are said to be *mutually independent* if for every subset

$$P\left(\bigcap_{j=1}^{m} A_{ij}\right) = \prod_{j=1}^{m} P(A_{ij})$$

These same events are *pairwise independent* if

$$P(A_i \cap A_j) = P(A_i)P(A_j)$$

Random Variables, Discrete Distributions

Definition: Let the sample space of an experiment be S. A *random variable* is a mapping from S to the real number line. If X is a random variable, then X associates with each $s \in S$.

Let X be a random variable on a sample space S, and let A be some subset of the real numbers. We then define $P(X \in A)$ by

$$P(X \in A) = P(\{s \in S : X(s) \in A\})$$

The probabilities so defined by all relevant subsets A is called the probability distribution of X.

Discrete vs. Continuous Random Variables: Since X is a mapping from S to the real numbers, the domain is S and the range we denote R_X.

If R_X is countable, X is a *discrete* random variable. If R_X is uncountable, X is a *continuous* random variable. When X is continuous, R_X is usually an interview or union of disjoint intervals. If S is countable X must be discrete, but if S is uncountable, X may be discrete or continuous.

The probability function (or probability mass function) of a discrete random variable is a function p_X given by

$$p_X(x) = P(X = x) \; \forall \; x \in \mathbb{R}$$

We then have

$$\sum_{i=1}^{\infty} p_X(x_i) = 1$$

For any subset A of real numbers, we may express $P(X \in A)$ as

$$P(X \in A) = \sum_{x \in A \cap R_X} p_X(x)$$

Bernoulli Distribution

The simplest discrete random variable takes only values 0 or 1. the probability mass function is

$$\begin{aligned} p_X(1) &= 0 \\ p_X(0) &= 1 - \theta \\ p_X(x) &= 0 \quad \text{otherwise} \end{aligned}$$

which can also be written as

$$\begin{aligned} p_X(x) &= \theta^x (1-\theta)^{1-x} \quad x = 0, 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Discrete Uniform Distribution

The discrete uniform frequency function p_X is defined by

$$\begin{aligned} p_X(x) &= 1/k \quad x = 0, 1...k \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Binomial Distribution

Observe n trials where the outcome is binary (success/failure). Probability of success is $0 \leq \theta \leq 1$ for each trial (each trial is independent). Define X to be the number of successes among n trials of a binomial experiment. The binomial probability mass function is

$$\begin{aligned} p_X(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad x = 0, 1...n \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Negative Binomial Distribution

Suppose we have a sequence of independent identical trials which can result in S or F, and the probability of S is θ for all trials. Continue the experiment until r successes have been observed (number of trials $= y + r$). If Y is the number of failures before r successes, we have the negative binomial random distribution pmf:

$$p_Y(y) = \binom{r-1+y}{r-1}\theta^r(1-\theta)^y \quad y = 0, 1...$$

Geometric Distribution

For the same scenario, if $r = 1$ and X is the number of failures until the first success, we have a geometric distribution:

$$p_X(x) = \theta(1-\theta)^x \quad x = 0, 1, 2...$$

Hypergeometric Distribution

If we sample *without replacement* from a finite population containing two types of items. Consider a population of N items containing M defective items and $N - M$ nondefective items. Randomly selecting $n \leq N$ items without replacement we define X to be the number of defectives in the sample. X has the pmf

$$\begin{aligned} p_X(x) &= \frac{\binom{n}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad x = M_1...M_2 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

where $M_1 = \max(0, n - (N - M))$ and $M_2 = (n, M)$.

Poisson Distribution

A random variable X, the number of successes occurring during a given time interval or specified region, is a Poisson random variable.

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

where λ is the rate for the given time or area. The pmf is

$$p_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, 2... \quad \lambda > 0$$

Continuous Random Variables

The pdf for a continuous random variable X over an interval (a, b) is

$$P(a \leq X \leq) = \int_a^b f_X(x)\,\mathrm{d}x$$

where any function f with the properties

- $f(x) \geq 0 \forall x$
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$

is a pdf. $f(x)$ is not itself a probability, but a probability density ('probability per unit x').

Uniform Distribution

For $L < R$ define f as

$$\begin{aligned} f(x) &= \frac{1}{R-L}, \quad L \leq x \leq R \\ &= 0, \quad \text{otherwise} \end{aligned}$$

The probability between two points is

$$\begin{aligned} P(x \leq X \leq x+\delta) &= \int_x^{x+\delta} f(t)\,\mathrm{d}t \\ &= \frac{1}{R-L}t\Big|_x^{x+\delta} \\ &= \frac{\delta}{R-L} \end{aligned}$$

Remark: Any nonnegative function g such that

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x < \infty$$

can be normalized into a pdf by multiplying by a constant.

Normal Distribution

Define

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

where the mean $\mu \in \mathbb{R}$ and standard deviation σ is any positive real number. X has the above density function by the notation

$$X \sim N(\mu, \sigma^2)$$

A *standard normal distribution* is given by the pdf ϕ and cdf Φ.

Gamma Distribution

The Gamma distribution is related to the Gamma function:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, \mathrm{d}x \,, \quad \alpha > 0$$

Some properties include:

- $\alpha > 1, \ \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
- if n is a positive integer, $\Gamma(n) = (n-1)!$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

The Gamma distribution is then defined by

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0,\infty)}(x)$$

where the shape parameter, α, and the inverse of the scale parameter, λ, can be any positive numbers.

The *kernel* of the distribution is the term $x^{\alpha-1}e^{-\lambda x}$.

Exponential Distribution

A special case of the gamma distribution in which $\alpha = 1$. The exponential density is given by

$$f(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$$

which has the property of *memorylessness*, meaning that a random variable X does not 'remember' previous events of the experiment.

Beta Distribution

Usually used to model random variables restricted to the interval (0,1). A random variable X with the pdf

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)$$

has the beta(a, b) distribution, which will be relevant for Bayesian statistics. The beta function is

$$\begin{aligned} B(a,b) &= \int_0^1 x^{a-1}(1-x)^{b-1} I_{(0,1)}(x) \, \mathrm{d}x \\ &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \end{aligned}$$

the pdf can also be written

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)$$

Definition: Any random variable, whether discrete or continuous, may be characterized by its *cumulative distribution function*, cdf, which is also called its distribution function. The cdf F_X of a random variable X is defined by

$$F_X(x) = P(X \leq x) \ \forall x$$

For a discrete X with pmf p_X we can say

$$F_X(x) = \sum_{\{k:k\leq x\}} p_X(k) \ \forall x$$

and for a continuous X with pdf f_X

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

For any cdf F_X:

- F_X is a nondecreasing function
- F_X is right continuous
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to\infty} F_X(x) = 1$

For a continuous random variable X, the pdf is

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

Quantile Functions

the quantile function provides a value such that the cdf equals a given probability from 0 to 1. Suppose F is 1-1. Then we define the quantile function Q (or the p^{th} quantile x_p) as

$$Q(p) = x_p = F^{-1}(p), \quad 0 < p < 1$$

where $Q(p) = x_p$ is the number such that $F(x_p) = p$. (Note: F can only be 1-1 when F has no flat regions except when 0 or 1.)

Functions of a Random Variable

If we are interested in some random variable $Y = h(X)$, where X is some other random variable, we can also talk about the pmf of Y as function of the random variable X.

Functions of a Discrete Random Variable

Consider the case where X is discrete and has pmf p_X, and we want to know the pmf of $Y = h(X)$. We then have

$$\begin{aligned} p_Y(y) &= P(Y = y) \\ &= P(h(X) = y) \\ &= \sum_{\{x:h(x)=y\}} p_X(x) \end{aligned}$$

Functions of a Continuous Random Variable

If X is a continuous random variable, consider $Y = h(X)$. The cdf of Y is

$$\begin{aligned} F_Y(y) &= P(h(x) \leq y) \\ &= \int_{\{x:h(x)\leq y\}} f_X(x)\,\mathrm{d}x \end{aligned}$$

where $f_X(x)$ is the pdf of X. Y can be continuous or discrete if X is continuous. If Y is continuous, then the pdf is the derivative of the cdf

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}$$

Consider X has pdf f_X and $Y = aX + b$, where $a > 0$. The pdf of Y is

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

The cdf is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right) \\ f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y\left(\frac{y-b}{a}\right) \\ &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \end{aligned}$$

This Y is a linear transformation of X.

Consider a change of variables as follows

$$f_X(x)\,\mathrm{d}x = f_Y(y)\,\mathrm{d}y$$

Then we have

$$\frac{\mathrm{d}x}{\mathrm{d}y} f_X(x) = f_Y(y)$$

where we interpret both $\frac{\mathrm{d}x}{\mathrm{d}y}$ and f_X as functions of y.

Joint Distributions

In many experiments there are several numbers associated with each possible outcome. For the moment, suppose there are only two of interest. So, we have two functions, X and Y, each of which maps points in the sample space into points on the real number line. The joint behavior of two random variables, X and Y, depends on their joint probability distribution, which specifies $P[(X, Y) \in B]$ for any subset B of $\mathbb{R}^2$.

As for single random variables, we can specify a joint distribution using the joint cumulative distribution function:

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

Discrete Bivariate Distributions

Consider the joint distribution of X and Y. in the case where both are discrete, the joint distribution is a frequency function p where

$$p_{X,Y}(x, y) = P(X = x, Y = y) \quad \forall (x, y)$$

For any two dimensional set A

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

Multinomial Probability Distribution

- Experiment consisting of n trials
- Each trial can result in one of k mutually exclusive categories
- Trials are independent
- Probability of i^{th} category is same for each trial: θ_i where $\sum_i \theta_i = 1$

The observable random variables are $\{X_i\}$ where X_i is the number of trials resulting in the i^{th} category, where $\sum_i X_i = n$. The random variables $\{X_i\}$ have the multinomial distribution

$$\begin{aligned} p(x_1, ..., x_k) &= P(X_1 = x_1, ..., X_k = x_k) \\ &= \frac{n!}{x_1! ... x_k!} \theta_1^{x_1} ... \theta_k^{x_k} \end{aligned}$$

where there are x successes and $n - x$ failures.

Marginal Distributions

Suppose X and Y are random variables having some joint distribution. When we obtain the probability distribution of, say, X from the joint distribution, the former distribution is referred to

as the *marginal distribution* of X. The marginal pmf of X is

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= \sum_y p_{X,Y}(x, y) \end{aligned}$$

The marginal pmf of Y can be defined similarly.

Continuous Joint Distributions

If X and Y are both absolutely continuous, the joint distribution is given by a pdf f. For $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int \int_A f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

As in the univariate case, any f such that

$$\int \int_A f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1$$

is a pdf.

Bivariate Distributions for Continuous Random Variables

The joint cdf of two continuous random variables (X, Y) is

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v) \, \mathrm{d}u \, \mathrm{d}v \end{aligned}$$

Marginal Distributions of Continuous Random Variables

Similarly to the discrete case, we have the marginal pdfs:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}y \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}x \end{aligned}$$

Bivariate Normal Distribution

(X, Y) have a bivariate normal distribution if they have the joint pdf

$$f_{X,Y}(x, y) = N \exp\left\{ \frac{-1}{2(1-\rho^2)\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho \frac{(x-\mu_X)}{\sigma_X} \frac{(y-\mu_Y)}{\sigma_Y} + \frac{(x-\mu_Y)^2}{\sigma_Y^2} \right]} \right\}$$

Where N is the normalization factor

$$N = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

and ρ is the correlation factor between X and Y. $\rho = 0 \implies$ independence.

Independence of Random Variables

We have a more complete definition of independence where random variables X and Y are independent if and only if

$$P[(X \in A) \cap (Y \in B)] = P(X \in A)P(Y \in B)] \quad \forall\ A, B \in \mathbb{R}$$

X and Y, discrete random variables with the joint pmf $p_{X,Y}$ and marginal pmfs p_X and p_Y, are independent if and only if

$$p_{X,Y} = p_X(x)p_Y(y) \quad \forall\ (x, y)$$

X and Y, continuous random variables with the joint pdf f and marginal densities f_X and f_Y, are independent if and only if

$$f_{X,Y} = f_X(x)f_Y(y) \quad \forall\ (x, y)$$

Conditional Distributions

In the discrete case, X, Y have the joint $P_{X,Y}$ and X has marginal pmf p_X. If $p_X > 0$, the conditional pmf of Y given $X = x$ is

$$\begin{aligned} p_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= \frac{P[(X = x) \cap (Y = y)]}{P(X = x)} \\ &= \frac{p_{X,Y}}{p_X} \end{aligned}$$

This can be rewritten:

$$\begin{aligned} p_{X,Y} &= p_{Y|X}(y|x)p_X \\ &= p_{X|Y}(x|y)p_Y \end{aligned}$$

We can also sum over all y to get

$$p_X(x) = \sum_y p_{Y|X}(y|x)p_Y(y)$$

The continuous case follows similarly:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}}{f_X} \\ f_{X|Y}(x|y) &= \frac{f_{X,Y}}{f_Y} \end{aligned}$$

We also have

$$\begin{aligned} f_{X,Y} &= f_{Y|X}(y|x)f_X \\ &= f_{X|Y}(x|y)f_Y \end{aligned}$$

$$\begin{aligned} f_Y &= \int_{-\infty}^{\infty} f_{X,Y}\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X\,\mathrm{d}x \end{aligned}$$

Independence via Conditional Distributions

Discrete random variables X and Y are independent if and only if

$$p_{X|Y}(x|y) = p_X(x) \quad \forall\ x$$

In the continuous case, X and Y are independent if and only if

$$f_{X|Y}(x|y) = f_X(x) \quad \forall\ x$$

Extrema and Order Statistics

If we have $X_1, ...X_n$ independent random variables with a cdf F and pdf f. The *order statistics* are values of $X_1, ...X_n$ arrange in increasing order, denoted $X_{(1)} < ... < X_{(n)}$. The cdf of the minimum $X_{(1)}$:

$$\begin{aligned} F_V(V) &= P(V \leq v) \\ &= 1 - P(V > v) \\ &= 1 - P(X_1 > v, ...X_n > v) \\ &= 1 - \prod_{i=1}^{n} P(X_i > v) \\ &= 1 - (1 - F(v))^n \end{aligned}$$

and find the pdf

$$\begin{aligned} f_V(v) &= F_V'(v) \\ &= \frac{\mathrm{d}}{\mathrm{d}v}[1 - (1 - F(v))^n] \\ &= n(1 - (1 - F(v))^{n-1}) f(v) \end{aligned}$$

Similarly, the cdf of the maximum, $U = X_{(n)}$, is

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= \prod_{i=1}^{n} P(X_i \leq u) \\ &= [F(u)]^n \end{aligned}$$

Convolutions

Sums of independent random variables X and Y, $Z = X + Y$:

In the discrete case:

$$\begin{aligned} p_Z(z) &= P(Z = z) \\ &= P(X + Y = z) \\ &= \sum_{\{(x,y):x+y=z\}} p_X p_Y \\ &= \sum_y p_X(z - y) p_Y(y) \\ &= \sum_x p_Y(z - x) p_X(x) \end{aligned}$$

In the continuous case:

$$\begin{aligned} F_Z(z) &= P(Z + Y \leq z) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y) \, \mathrm{d}x \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_Y(y) \left[\int_{-\infty}^{z-y} f_X(x) \, \mathrm{d}x \right] \mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_Y(y) F_X(z - y) \, \mathrm{d}y \end{aligned}$$

and we find the density by differentiating

$$\begin{aligned} f_Z &= \frac{\mathrm{d}F_Z(z)}{\mathrm{d}z} \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, \mathrm{d}x \end{aligned}$$

which is called the convolution of f_X and f_Y.

Expectations, Variance, Covariance and Correlation

The expected value, or mean, of a discrete random variable X with pmf p_X is

$$\begin{aligned} E(X) &= \mu_X \\ &= \sum_x x p_X(x) \end{aligned}$$

over all x such that $p_X(x) > 0$.

In the continuous case:

$$\begin{aligned} E(X) &= \mu_X \\ &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \end{aligned}$$

Expectations of Functions of a Random Variable

Consider a random variable $Y = g(X)$, and p_Y or f_Y be the pmf or pdf of Y, respectively. The expectation value of the random variably Y is

$$E(Y) = \begin{cases} \sum_y y p_Y(y) & Y \text{ is discrete} \\ \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y & Y \text{ is continuous} \end{cases}$$

which can also be computed

$$E(g(X)) = \begin{cases} \sum_x g(x) p_X(x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x & x \text{ is continuous} \end{cases}$$

Properties of Expectations

Expectation of a linear function of a random variable: Let $Y = aX + b$, where a and b are constants. If $E(X)$ exists, then

$$E(Y) = aE(x) + b$$

Expectation of a sum of random variables: Let $Y = a + b_1 X_1 + ...b_n X_n$. If $E(X_i)$ exists then

$$E(Y) = a + b_1 E(X_1) + ... + b_n E(X_n)$$

Expectation of a product of independent random variables: Suppose $X_1 ... X_n$ are independent random variables, and let $h_1, ..., h_n$ be functions such that $E[h_i(X_i)]$ exists,

$$E\left[\prod_{i=1}^{n} h_i(X_i)\right] = \prod_{i=1}^{n} E[h_i(X_i)]$$

Monotonicity of Expectation: Let X be a random variable where $P[X \geq 0] = 1$. Then

$$\begin{aligned} E(X) &\geq 0 \\ &> 0 \iff P[X - 0] < 1 \end{aligned}$$

If X and Y are random variables such that $X \leq Y$, then

$$E(X) \leq E(Y)$$

Variance of a Random Variable

Suppose X is a random variable such that $E(X) = mu_X$ exists and is finite. The *variance* of X is defined to be

$$\sigma_X^2 = E[(x - \mu_X)^2]$$

and the *standard deviation* of a random variable is defined by

$$\begin{aligned}\sigma_X &= \sqrt{\sigma_X^2} \\ &= \sqrt{E[(x - \mu_X)^2]}\end{aligned}$$

Some properties of variance:

- The variance as a function of expectations of the random variable can be written

$$\begin{aligned}\sigma_X^2 &= E(X^2) + \mu_X^2 \\ &= E(X^2) - [E(X)]^2\end{aligned}$$

- $\sigma_X^2 = 0 \iff \exists\, c:\ P(X = c) = 1$
- If a and b are constants, $\text{Var}(aX + b) = a^2\text{Var}(X)$, and the standard deviation is $|a|\sigma_X$. If $a = 1$, we have $\text{Var}(X + b) = \text{Var}(X)$, so shifting a distribution by some b does not change the variance.
- For $X_1, ..., X_n$ independent random variables, $a_1, ..., a_n$ constants,

$$\text{Var}(a_1X_1 + ...a_nX_n) = \sum_{i=1}^{n} a_1^2\text{Var}(X_i)$$

  If all $a_i = 1$, we have the variance of a sum of independent random variables is the sum of the variances, which is not generally true for random variables which are not independent.

Covariance and Correlation

The covariance of X and Y is denoted $\text{Cov}(X, Y)$ and defined by

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_x)(Y - \mu_Y)] \\ &= \text{Cov}(Y, X)\end{aligned}$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$. The covariance measures the tendency of X and Y to be on the same (or opposite) sides of their respective means.

The correlation between X and Y is a scaled version of the covariance. It does not depend on the measurement units of X and Y. The correlation is denoted

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}$$

This has the property $-1 \leq \text{Corr}(X, Y) \leq 1$. Some other properties of covariance and correlation:

- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
- if X and Y are independent, $0 < \sigma_X < \infty$ and $0 < \sigma_Y < \infty$, then

$$\text{Cov}(X, Y) = 0$$
$$\text{Corr}(X, Y) = 0$$

- $\text{Cov}(X, X) = \text{Var}(X)$
- For some constants a, b, where $a \neq 0$, if $Y = aX + b$

$$\text{Corr}(X, Y) = \begin{cases} 1 & a > 0 \\ -1 & a < 0 \end{cases}$$

- 'X and Y are independent' $\implies$ $\text{Cov}(X, Y) = 0$. The converse statement is not necessarily true.

Variance of a Linear Combination

Suppose X and Y are jointly distributed random variables and define $Z = aX + bY$, where a, b are constants.

$$E(Z) = a\mu_X + b\mu_Y$$
$$\text{Var}(Z) = a^2 \text{ Var}(X) + b^2 \text{ Var}(Y) + 2ab \text{ Cov}(X, Y)$$

If the covariance is zero

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

For a linear combination of n random variables

$$U = a + \sum_{i=1}^{n} b_i X_i$$
$$\text{Var}(U) = \sum_{i=1}^{n} b_i^2 \text{Var}(X_i) + 2 \sum_{i<j}^{n} b_i b_j \text{Cov}(X_i, X_j)$$

Moments and the Moment-Generating Function

Definition: The *moments* of a random variable X are (when they exist) the expectations of powers of X

$$E(X^k), \quad k = 1, 2, ...$$

Definition: The *central moments* of X are

$$E[(X - \mu)^k], \quad k = 2, 3, ...$$

The second central moment is

$$\begin{aligned} E[(X-\mu)^2] &= E(X^2) - \mu^2 \\ &= \text{Var}(X) \end{aligned}$$

Consider the following function of s

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

If there exists a positive number s_0 such that the last expectation exists for all $|s| < s_0$, then $M_X(s),\ |s| < s_0$ is called the *moment generating function* or mgf of X. When the mgf of X exists, then all the moments of X exist and are finite. Furthermore, we may find the moments of X from M_X as

$$E(X^k) = \left.\frac{\mathrm{d}^k M_X(s)}{\mathrm{d}s^k}\right|_{s=0}$$

Where, for example, $M'_X(0) = E(X)$ and $M''_X(0) = E(X^2)$. Also note that $M_X(0) = 1$ for any mgf.

Properties of mgfs:

- $M_X(0) = 1$
- The mgf of $Y = aX + b$ is

$$\begin{aligned} M_Y(s) &= E(e^{sY}) \\ &= e^{bs} M_X(as) \end{aligned}$$

- For $X_i ... X_n$ independent random variables with respective mdfs $M_z ... M_z$,

$$M(S) = \prod_{i=1}^{n} M_i(s)$$

- The distribution of X is the same as that of Y if and only if $M_X = M_Y$

Conditional Expectation and Prediction

A conditional expectation is an expectation defined with respect to a conditional distribution.

$$E[h(Y)|X = x] = \begin{cases} \sum_y h(y) p_{Y|X}(y|x) & \text{for discrete } Y \\ \int_{-\infty}^{\infty} h(y) f_{Y|X}(y|x)\,\mathrm{d}y & \text{for discrete } Y \end{cases}$$

$\mu(x) = E(Y|X = x)$ is known as the regression function of Y on x. The regression function us often used to predict Y for a given $X = x$. Some properties of conditional expectation and conditional variance:

- $\text{Var}(Y|X = x) = E[(Y - \mu(x))^2]$

Theorems:

- Iterated expectation: $E[E(Y|X)] = E(Y)$
- Variance partition: $\text{Var}(Y) = \text{Var}[E(Y|X)] + E[\text{Var}(Y|X)]$

Markov's Inequality: Let X be a random variable such that $P(X \geq 0) = 1$. Then, for every positive number a

$$P(X \geq a) \leq \frac{E(X)}{a}$$

is useful for large $a : a > E(X)$.

Consider the indicator function $I_{X \geq a}$. We have

$$I \leq \frac{x}{a} E(I_{X \geq a}) \leq \frac{E(X)}{a}$$

Chebyshev's Inequality: Suppose X is a random variable with finite variance. For each $a > 0$:

$$P(|X - \mu_X| \geq a) \leq \frac{\sigma_X^2}{a^2}$$

Random Sampling

$X_1, ...X_n$ is a *random sample* of a pdf/pmf f_θ. θ represents an unknown parameter that specifies the specific distribution within a family of distribution. We have the maximum likelihood estimator:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n X_i}$$

We can use mgfs to obtain the distribution of the sum of the random variables in a random sample $Y = \sum_{i=1}^n X_i$.

$$\begin{aligned} M_Y(s) &= \prod_{i=1}^n E(e^{sX_i}) \\ &= \left(\frac{\lambda}{\lambda - s}\right)^n \end{aligned}$$

which is the mgf of the gamma(n, λ) distribution.

A few important laws and theorems:

- *The Weak Law of Large Numbers* - tells us that a sample mean will be very close (with high probability) to the population mean for a large enough sample.
- *The Central Limit Theorem* enables us to use the normal distribution to approximate the sampling distribution of the sample mean for large samples.

Convergence in Probability and the Law of Large Numbers

Estimating Suppose $X_1, ...X_n$ is a random sample from a distribution with mean μ_X and variance σ_X^2. Consider the problem of estimating the mean μ_X using the sample mean $\bar{X}_n$

$$\begin{aligned} \bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i & E(\bar{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{n\mu_X}{n} \\ \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{\sigma_X^2}{n} \end{aligned}$$

Chebyshev's inequality gives

$$P(\left|\bar{X}_n - \mu_X\right| \geq a) \leq \frac{\sigma_X^2}{na^2}$$

Assuming we know the variance, we can use this to determine a sample size n that guarantees

$$P(\left|\bar{X}_n - \mu_X\right| < a_0) \geq p$$

set

$$\begin{aligned} 1 - p &= \frac{\sigma_X^2}{na^2} \\ \implies n &= \frac{\sigma^2}{(1-p)a_0^2} \end{aligned}$$

Weak Law of Large Numbers

$$0 \leq \lim_{n\to\infty} P(\left|\bar{X}_n - \mu_X\right| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

Convergence of Probability

Definition: Let $Y_1, Y_2...$ be a sequence of random variables. $\{Y_i\}$ *converges in probability* to the constant b if for every $\epsilon > 0$.

$$\lim_{n\to\infty} P(|Y_n - b| \geq \epsilon) = 0$$

We write $Y \xrightarrow{P} b$.

Monte Carlo Approximation of an Integral

Consider an integral I which we want to approximate numerically. If the integral can be represented by the expectation of a function h of a random variable X from a given distribution $f_X(x)$ as

$$I = \int_{R_X} h(X) f_X(x)\,\mathrm{d}x$$

We can approximate the integral by

1. select a large positive integer n
2. generate a random sample from the distribution with pdf f_X
3. set $T_i = h(X_i)$
4. Estimate I by $M_n = (T_i...T_n)/n$

The Weak Law of Large Numbers implies

Convergence in Distribution and the Central Limit Theorem

For a special case of a random sample from a $N(\mu, \sigma^2)$ distribution, with sample mean $\bar{X}_n$, we know

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

so

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

Standardizing the sample mean gives

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

We derive *confidence intervals* using

$$P\left(-Z_{(1+\gamma)/2} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < Z_{(1+\gamma)/2}\right) = \Phi(-Z_{(1+\gamma)/2}) - \Phi(Z_{(1+\gamma)/2})$$
$$= \gamma$$

Definition: Let $X_1, X_2...$ be a sequence of random variables with cdfs $F_1, F_2....$ We say X_n *converges to* X *in distribution* if

$$\lim_{n\to\infty} F_n(X) = F(X)$$
$$X_n \xrightarrow{D} X$$

Central Limit Theorem

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

The sequence $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},\ n = 1, 2, ...$ converges in distribution to Z, a standard normal random variable. We can also express the CLT in terms of the sum $S_n = X_1 + ... + X_n$, where

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z)$$

For a set of independent random variables $X_1, ...X_n$ from random distributions $X_i \sim N(\mu_i, \sigma_i^2),\ i = 1, ...n$. For a linear combination $Y = b + a_1X_1 + ...a_nX_n$, we have

$$Y \sim N\left(b + \sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2, \right)$$

If the random sample is from $N(\mu, \sigma^2)$, then the sample mean is

$$\begin{aligned}\bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &\sim N(\mu, \sigma^2/n)\end{aligned}$$

Considering $U = a + \sum_{i=1}^{n} b_iX_i$ and $V = c + \sum_{j=1}^{n} d_jX_j$ gives...

The Chi-Squared Distribution

The chi-squared parameter with n degrees of freedom is a gamma distribution with $\alpha = n/2,\ \lambda = 1/2$. It has the pdf

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)}x^{(n/2)-1}e^{-x/2}, \quad x > 0$$

For random variable X from the chi-squared distribution, we write $X \sim \chi^2(n)$. If $Z \sim N(0, 1)$, then $Z^2 \sim \chi^2(1)$. For a random sample $Z_1, ...Z_n$ from the $N(0, 1)$, then

$$Z_1^2 + ... + Z_n^2 = \chi^2(n)$$

The t Distribution Consider the random variables $Z \sim N(0, 1)$ and $U \sim \chi^2(n)$, then the following random variable has a t distribution $t(n)$

$$T = \frac{Z}{\left(\frac{U}{n}\right)^{1/2}}$$

so the pdf is

$$f(t) - \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

- Each $t(n)$ pdf is symmetric about zero
- For $n = 1$ the $t(1)$ distribution is the same as the Cauchy distribution
- Each $t(n)$ pdf has a smaller peak and thicker tails than the standard normal pdf
- In the limit $n \to \infty$, the $t(n)$ pdf approaches the standard normal

- For $n > 1,\ E(T) = 0$
- For $n > 2,\ \text{Var}(T) = n/(n-1)$

The F Distribution Consider two independent random variables $U \sim \chi^2(m)$, $V \sim \chi^2(n)$, the following random variable follows the F distribution $F(m,n)$ with degrees of freedom m, n

$$W = \frac{U/m}{V/n}$$

The pdf of the $F(m,n)$ distribution

$$f(w) - \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2}, \quad w > 0$$

For $n > 2,\ E(W) = n/(n-2)$.

Sampling Distribution of Sample Mean and Sample Variance

Consider a random sample $X_1, ...X_n$ from $N(\mu, \sigma^2)$. The population mean and variance are

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

The sample mean and sample variance satisfy

$$\bar{X} = N(\mu, \sigma^2/n)$$
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)$$

Independence of the sample mean and sample variance characterizes the normal distribution. If the sample mean and sample variance are independent, then the random sample must be from the normal distribution.

Let

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$
$$U = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

so the random variable T

$$T = \frac{Z}{\left(\frac{U}{n-1}\right)^{1/2}}$$

follows $T \sim t(n-1)$. In terms of the random sample we have

$$T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}$$

where

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

Probability Models and Statistical Models, Statistical Inference

---

- Collect some data (a sample of the population)
- We want to know certain aspects of the population
- The data collected contain information concerning the population
- Summarize the properties of the population using probability models
- *Statistical inference* involves using the collected data to make inferences about the population

Statistical inference involves *uncertainty*. Probability quantifies the amount of uncertainty in an inference.

- We have a random sample $X_1, ...X_n$ from some distribution f.
- Assume the form of f is known except for some parameters $\theta_1, ...\theta_k$.
- Use the random sample to draw conclusions about the unknowns.

The pdf or pdf f or p is referred to as the *population*.

$$f(x) = f_{\theta_1,...\theta_k}(x)$$

The *parameter space* is the collection of all possible parameter values, denoted Ω. For the normal distribution example:

$$\Omega = \{\{\theta_1, \theta_2\} : -\infty < \theta_1 < \infty,\ \theta_2 > 0\}$$

A function of the random sample which does not depend on any unknown parameters is called a *statistic*. Inferences must depend only on the statistic, not the unknown parameters.

Types of Inference

- Point Estimation: Computing a single value from the data that is thought to be near the true value of the parameter
- Interval Estimation: Computing from the data a range of plausible values for the unknown parameter

- Hypothesis Testing: Using the data to choose between two statements concerning the unknown parameter
- Prediction: Using the data to predict future values coming from the population
- Goodness of Fit: Using the data to assess whether the assumed distribution is a reasonable model

The Likelihood Function

Suppose we observe data s and that the pmf is p_θ. The likelihood function is defined on the parameter space Ω by

$$L(\theta|s) = p_\theta(s) \quad \theta \in \Omega$$

The likelihood is the probability of observing s when the true value of the parameter is θ, which induces an ordering on Ω in that we believe that θ_1 is more probable that θ_2 if

$$p_(\theta_1) > p_\theta(s)$$
$$L(\theta_1|s) > L(\theta_2|s)$$

The likelihood function for a simple random sample from pdf or pmf f_θ with the joint pdf or pmf

$$f_\theta(x_1, ...x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

is

$$L(\theta|x_1, ...x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

The *log-likelihood function* is defined

$$\begin{aligned} l(\theta|s) &\equiv \log(L(\theta|s)) \\ &= \prod_{i=1}^{n} \log(f_\theta(x_i)) \end{aligned}$$

Sufficient Statistics

A statistic $T(s)$ is a *sufficient statistic* for θ if, whenever $T(s_1) = T(s_2)$:

$$L(\theta|s_1) = c(s_1, s_2)L(\theta|s_2)$$

Factorization Theorem

$T(s)$ is a sufficient statistic for the model if the density factors

$$f_\theta(s) = h(s)g_\theta(T(s))$$

where g_θ and h are nonnegative and h is not dependent on θ.

Maximum Likelihood Estimates

The *maximum likelihood estimates* are values $\hat{\theta}$ such that

$$L(\hat{\theta}(s)|s) > L(\theta|s) \quad \forall\ \theta \in \Omega$$

Score Function and Calculation of the MLE

The *score function* is the derivative of the log-likelihood with respect to θ.

$$S(\theta|s) = \frac{\partial l(\theta|s)}{\partial s}$$

We solve for the MLE by finding the solution(s) of

$$S(\theta|s) = 0$$
$$\left.\frac{\partial^2 l(\theta|s)}{\partial s^2}\right|_{\theta=\hat{\theta}} < 0$$

Invariance Property of MLEs

Given a random sample from some distribution with pdf or pmf f_θ and mle $\hat{\theta}$, we can alternatively parameterize $\tau = \psi(\theta)$ where ψ is a one-to one function of θ. the *plug-in estimate* of τ is $\hat{\tau} = \psi(\hat{\theta})$. The invariance property of mles says:

If $\hat{\theta}$ is the mle of θ and ψ is a one-to-one function of θ, then $\psi(\hat{\theta})$ is the mle of $\psi(\theta)$. This may be badly-behaved if ψ is not one-to-one.

Methods of Moment Estimators, Properties of Estimators, Bias

The first population moment of a random variable X is

$$\mu_X = E_\theta(X)$$

Now we introduce *sample moments*. Given a random sample, the first sample moment is $m_1 = \bar{X}$. With the method of moments, we express the population mean as a function of the unknown parameter θ, solve for θ, and substitute the sample mean for the population mean.

Definition: The *bias* of a general estimator T of the parameterization $\psi(\theta)$ is defined

$$\text{Bias}_\theta(T) = E_\theta(T) - \psi(\theta)$$

If $\text{Bias}_\theta(T) = 0\ \forall\ \theta \in \Omega$, we say that T is an *unbiased estimator* of $\psi(\theta)$.

Definition: The most commonly reported measure of the variation in the estimator T is the *standard error*:

$$\text{SE}_\theta(T) = \sqrt{\text{Var}_\theta(T)}$$

Definition: The *mean squared error* judges how well $\hat{\theta}$ estimates θ. The MSE of an estimator T of $\psi(\theta)$ is

$$\begin{aligned}\text{MSE}_\theta(T) &= E_\theta((T - \psi(\theta))^2) \\ &= \text{Var}_\theta(T) + [\text{Bias}_\theta(T)]^2\end{aligned}$$

Consistency for Estimators

Let T_n be an estimator of $\psi(\theta)$ for a random sample from $f_\theta(x)$. The sequence of estimators $\{T_n, n = 1, 2, ...\}$ is *consistent for* $\psi(\theta)$ if

$$T_n \xrightarrow{P} \psi(\theta)$$

From the weak law of large numbers, the sample moments converge in probability to the population moments:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X^k)$$

A convenient method of showing that an estimator is consistent is to show that the MSE goes to zero at large n.

$$\begin{aligned} \mathrm{Var}_\theta(T_n) &\to 0 \quad n \to \infty \\ \mathrm{Bias}_\theta(T_n) &\to 0 \quad n \to \infty \end{aligned}$$

Asymptotic Distribution of Estimators

For a random sample of a distribution with finite variance, for each $z \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

which implies

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \ \sim \ N(0,1)$$

The Delta Method

Suppose we know

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, V(\theta))$$

and we are interested in the asymptotic distribution of the estimator $g(T_n)$ of $g(\theta)$. If g is continuous with a sufficient number of derivatives, Taylor expand:

$$g(t) = g(\theta) + (t - \theta)g'(\theta) + ...$$

Taylor expand to T_n and ignore higher order terms

$$\begin{aligned} g(T_n) &= g(\theta) + (T_n - \theta)g'(\theta) \\ \sqrt{n}(g(T_n) - g(\theta)) &\xrightarrow{D} N(0, [g'(\theta)]^2 V(\theta)) \end{aligned}$$

Bootsrapping

We can approximate the standard error of an estimator computationally by *bootstrapping*. Suppose we have $\hat{\theta} = S(X_1, ..., X_n)$ for some statistic S.

- if we knew the true value of the unknown parameter, θ_0, we could generate $x_i, ..., x_n$ from $f_{\theta_0}(x)$ and compute $s_1 = s(x_1, ...x_n$.
- Repeat this a large number, B, times to obtain a random sample of values of S, $(s_1, ...s_B)$.
- estimate the sampling distribution of S
- Replace θ_0 by $\hat{\theta}$

Large Sample Theory for Maximum Likelihood Estimators

Some large-sample properties of MLEs. For a random sample from a population with a given pdf or pmf:

- The density is a snooth function of the unknown parameter θ
- The support of the distribution does not depend on the unknown parameter θ
- the parameter space Ω satisfies certain conditions
- The variance of the derivative of the log of the density with respect to the unknown parameter is finite.

Some important properties of the MLE include

- The MLE $\hat{\theta}_n$ is consistent
- The MLE is asymptotically normal

Definition: Fisher's information in θ is based on the observation X is defined as

$$\begin{aligned} I(\theta) &= E_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] \\ &= -E_\theta \left[\left(\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right) \right] \end{aligned}$$

Confidence Intervals

For some random sample from a normal distribution with known variance, construct a random interval which will contain the mean μ with a specified probability γ.

$$P[l(X_1, ..., X_n) \leq u(X_1, ..., X_n)] = \gamma$$

We can use the standard normal distribution as:

$$P(-c < Z < c) = \Phi(c) - \Phi(-c) = \gamma$$

$$Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

$-c < Z < c$ implies:

$$P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

A general definition of a confidence interval $C(X_1, ...X_n) = (l(X_1, ...X_n), u(X_1, ...X_n))$ is such that

$$P(\psi(\theta) \in C(X_1, ...X_n)) \geq \gamma$$

Pivots Suppose we have a random sample from distribution $f_\theta(x)$. A pivot W is a random variable whose distribution does not depend on θ.

$$P(a < W < b) = \gamma$$

The confidence interval for θ can be written

$$C(x_1, ..., x_n) = \{\theta : \ a < W < b\}$$

We can find approximate confidence intervals based on the MLEs

$$\hat{\theta} \approx N(\theta_0, \hat{V}_n)$$

We can form the approximate pivot

$$\frac{\hat{\theta} - \theta}{\hat{V}_n} \approx N(0, 1)$$

An approximate 100% confidence interval for θ is

$$\hat{\theta} \pm Z_{(1+\gamma)/2} ASE(\hat{\theta})$$

This forms the **Wald confidence interval**.

Bayesian Approach to Parameter Estimation

In a Bayesian approach, we suppose that the unknown parameter θ is a random variable with a *prior distribution* $\pi(\theta)$. For a given value of the unknown parameter, the data S have a pdf or pmf $f_\theta(s)$. The joint distribution is

$$f(s, \theta) = f_\theta(s)\pi(\theta)$$

The marginal distribution of S is

$$m(s) = \int_\Omega f_\theta(s)\pi(\theta)\, \mathrm{d}\theta$$

For the purposes of inference, we are interested in the conditional *posterior* distribution of θ given $S = s$.

$$\begin{aligned} \pi(\theta|s) &= \frac{f(s, \theta)}{m(s)} \\ &= \frac{f_\theta(s)\pi(\theta)}{\int_\Omega f_\theta(s)\pi(\theta)\, \mathrm{d}\theta} \end{aligned}$$

The posterior represents the knowledge of the statistician arising from prior information and from the data. We can say that the posterior density is proportional to the likelihood times the prior density:

$$\pi(\theta|s) \propto f_\theta(s)\pi(\theta)$$

- The mode of the posterior distribution represents the most likely value of θ given $S = s$
- The mean (or median) of the posterior distribution represents the mean (or median) value of θ given $S = s$
- The The standard deviation and variance of the posterior distribution describe the variability of θ given $S = s$

Choice of Prior Distribution

A *noninformative prior* treats all possible values of the parameter equally. An *improper prior* is a weight function which does not have a finite integral. A family Π of distributions is a conjugate family for $\mathcal{F} = \{f(x|\theta) : \theta \in \Omega\}$ if when whe choose a prior distribution $\pi \in \Pi$, the posterior is also in π.

Bayes Credible Interval

A γ *Bayes credible interval* for the parameter $\psi(\theta)$ is an interval $C(s) = [l(s), u(s)]$ such that

$$P[\psi(\theta) \in C(s)|S = s] = \gamma$$

A *highest posterior density interval* satisfies

$$\int_{l(s)}^{u(s)} \pi(\theta|s)\,\mathrm{d}\theta = \gamma$$

For a random sample from a normal population with mean μ and variance σ^2, we can define the precision $\xi = \frac{1}{\sigma^2}$.The pdf becomes

$$f(x|\mu, \xi) = \left(\frac{\xi}{2\pi}\right)^{1/2} e^{-\frac{1}{2}\xi(x-\mu)^2}$$

In the case where the precision is known $\xi = \xi_0$, the conjugate prior is $N(\mu_0, \xi_{prior}^{-}1)$. The posterior has a mean

$$\mu_{post} = \bar{x}\frac{n\xi_0}{n\xi_0 + \xi_{prior}} + \mu_0\frac{n\xi_{prior}}{n\xi_0 + \xi_{prior}}$$

which is a weighted average of the prior mean μ_0 and the sample mean (MLE) $\bar{x}$.

Testing Hypotheses

The *null hypothesis* $H_0 : \psi(\theta) = \psi_0$ is tested to ascertain whether a hypothesized value of a characteristic $\psi(\theta)$ of the population is consistent with the observed data s. A *hypothesis test* provides a measure of how unlikely the observed data appear under the assumption that the null

hypothesis is true. The *p-value* indicates the evidence for the null hypothesis, where a small p-value indicates that the null hypothesis should be rejected.

Neyman-Pearson Paradigm for Hypothesis Testing

The Neyman-Pearson approach does not use prior probabilities, but concentrates on the two error probabilities. We observe the data and conclude that we *reject* the null hypothesis or *fail to reject* the null hypothesis. Possible errors:

- Type I error: Reject H_0 when H_0 is true. Probability α (level of significance)
- Type II error: Do not reject H_0 when H_0 is false. Probability $1-\beta$.The *power* is the probability that H_0 is rejected when it is false and is equal to β.

Neyman-Pearson Lemma

Suppose that $H_0 : \theta = \theta_0$ and $H_a : \theta = \theta_1$ are simple hypotheses. Consider the test that rejects H_0 whenever the likelihood ratio $f_{\theta_1}/f_{\theta_0} > c_0$ and suppose it has size α. Then any other test which has size $\leq \alpha$ has power less than or equal to that of the likelihood ratio test.

P-Value

The P-value shows the observed significance. as the smallest level of significance at which H_0 can be rejected using a rejection region of the given form. Another definition of P-value is the probability of a result at least as extreme as the observed test statistic when H_0 is true.

Remarks:

- The P-value is a statistic calculated from the data
- Under fairly general conditions when the test statistic has a continuous distribution, the P-value is uniformly distributed on [0,1]

Generalized Likelihood Ratio Tests

For a random sample we consider the hypotheses $H_0 : \ \mu = \mu_0$ and $H_a : \ \mu \neq \mu_0$. IF we are interested in a specific alternative hypothesis, such as $H_a : \ \mu = \mu'$, The NP lemma implies that we should use a likelihood ratio as our test statistic. Construct a generalized LR test for a normal mean using

- Write the LR statistic for testing two hypotheses
- Substitute the MLE for the mean under the alternative
- Rewrite the test statistic to make a new test statistic for a known distribution
- Find the rejection region for the new statistic

Under the conditions for the asymptotic normality of the MLE, the null distribution of $2\log$ LR converges to a chi-squared distribution with degrees of freedom $df = \dim(\Omega) - \dim(H_0)$ as n tends to infinity.

Forming Confidence Intervals

Generalized LR test rejects when

$$\begin{aligned}\frac{n(\bar{x}-\mu_0)^2}{\sigma_0^2} &> \chi^2_{1-\alpha}(1)\\ &> Z^2_{1-\alpha/2}\end{aligned}$$

When σ_0^2 is unknown, the rejection region is

$$\frac{n(\bar{x}-\mu_0)^2}{S^2} > t^2_{1-\alpha}(n-1)$$