C&EE 110:
Introduction to Statistics and Probability

Einar Balan

Contents

1 | Sets and Probablity Theory

1.1 Probabilistic Sets

A random event, E, has more than 1 possible outcome in the sample space S. S is the collection of all possible event outcomes. We know that E ⊂ S.

Ex) Number in dice roll

$$S = \{1, 2, 3, 4, 5, 6\}$$
$$E_{odd} = \{1, 3, 5\}$$
$$E_{>3} = \{4, 5, 6\}$$

Operations

We can apply several operations to our sets.

1. Union, denoted $E_1 \cup E_2$
2. Intersection, denoted $E_1 \cap E_2$ or $E_1 E_2$

Consider E_{odd} and $E_{>3}$ above.

$$E_{odd} \cup E_{>3} = \{1, 3, 4, 5, 6\}$$
$$E_{odd} \cap E_{>3} = \{5\}$$

These operations are commutative, associate, and distributive. Intersection has precedence over union.

Special Events

- S is the event the spans the entire sample space
- $\varnothing$ is the null event, it has no outcomes
- if E_1 and E_2 are mutually exclusive, $E_1 E_2 = \varnothing$
- if E_1 and E_2 are collectively exhaustive, $E_1 \cup E_2 = S$
- $\overline{E_1} = S - E_1$, the complement[1] of E_1

[1] Demorgan's Laws hold

Frequentist Probability (Natural Variation)

The probability of occurrence of E is the relative frequency of observations of E ina large number of repeated experiments. Put more formally below,

$$P(E) = \lim_{N\to\infty} \frac{n}{N}, \text{ where n = occurences of E in N observations in S}$$

Bayesian Probability (Incomplete Knowledge)

The probability of an event E represents analysts' degree of belief that E will occur.

Frequentist Probability	Bayesian Probability
probability of expecting a ground shaking intensity of 1g in next 100 years	probability of finding water on new planet
max wind speed in a year	probability that a buidling will collapse under ground shaking intensity of 1g
live load on a building	election results
*based on previous observations	*not based on previous observations
*cannot be reduced through more measurement	*can be reduced if more observations/measurements applied

1.2 Axioms

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B)$, s.t. $AB = \varnothing$

* these axioms are consistent with Frequentist probability

We can derive several rules from these axioms.

1. $P(\overline{E}) = 1 - P(E)$
2. $P(\varnothing) = 0$
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1E_2)$
    - if E_1 and E_2 are mutually exclusive, then we double count their intersection when using the 3rd axiom; subtracting it leads to the correct value
    - what if we have $>$ 2 events? Inclusion/Exclusion rule
    - $P(E_1 \cup E_2 \cup ... \cup E_n) =$

$$\sum_{i=1}^{n} P(E_i) - \sum_{i=1}^{n}\sum_{j=1}^{i-1} P(E_iE_j) + \sum_{i=1}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} P(E_iE_jE_k) + ... + (-1)^{n-1}P(E_1E_2...E_n)$$

Conditional Probability

We may want to determine the probability of an event given another event is guaranteed to occur. This is denoted $P(E_1|E_2)$, which is read as E_1 given E_2. It essentially redefines the sample space to be E_2.

$$P(E_1|E_2) = \begin{cases} \frac{P(E_1E_2)}{P(E_2)} & P(E_2) > 0 \\ 0 & P(E_2) = 0 \end{cases} \tag{1.1}$$

From this equation, it follows that

$$P(E_1E_2) = P(E_1|E_2)P(E_2)$$

This holds in general for n events.

$$P(E_1E_2E_3) = P(E_1|E_2E_3)P(E_2E_3) = P(E_1|E_2E_3)P(E_2|E_3)P(E_3)$$

Ex) Applying conditions to operations

$$P(E_1 \cup E_2|E_3) = P(E_1|E_3) + P(E_2|E_3) - P(E_1E_2|E_3)$$

$$P(E_1E_2|E_3) = P(E_1|E_2|E_3)P(E_2|E_3), \text{ which follows from 1.1}$$

Independence

Two events are indpendent iff $P(E_1|E_2) = P(E_1)$

We have mutual independence if $P(E_1E_2 \ldots E_n) = P(E_1)P(E_2)\ldots P(E_n)$.

Theorem of Total Probability

Consider an event A and a set of of mutually exclusive and collectively exhaustive events $E_1, E_2, \ldots, E_3$.

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i) \tag{1.2}$$

Bayes' Rule

Consider an event A and a set of of mutually exclusive and collectively exhaustive events $E_1, E_2, \ldots, E_3$ in S.

$$P(AE_j) = P(E_j|A)P(A) = P(A|E_j)P(E_j)$$

$$P(E_j|A) = \frac{P(A|E_j)P(E_j)}{P(A)}$$

$$P(E_j|A) = \frac{P(A|E_j)P(E_j)}{\sum_{i=1}^{n} P(A|E_i)P(E_i)}$$

where equation 1.2 is used to subtitute P(A)

2 | Random Variables

A **random variable** is a variable whose specific value cannot be predicted with certainty before an experiment. They take on a numerical value for each possible event in the sample space.

Ex) Random variables are easy to define. For example,

- X = magnitude of a future earthquake
- Y = yield stress of a material
- Z = peak wind pressure during a given year

For a random variable X, its outcomes are denoted $x_1, x_2, \dots, x_n$. For an outcome x_i, we denote the probability of that outcome as $P(X = x_i)$.

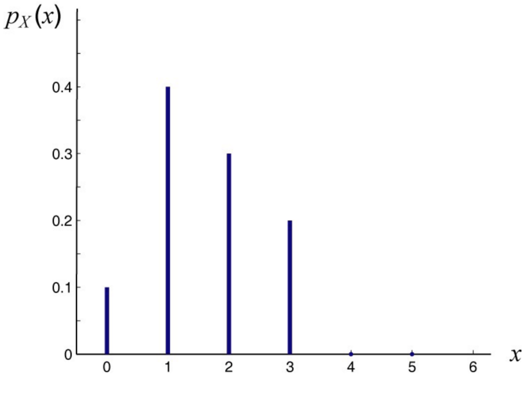
2.1 Discrete Random Variables

A random variable is called **discrete** if the number of outcomes is countable. For example, for X = the number of cars on a bridge at a certain time, X is discrete.

Distributions of discrete random variables can be quantified in 2 ways.

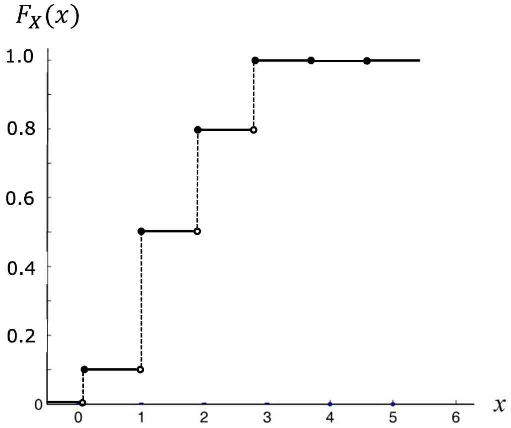
1. Probability Mass Function

$$p_X(x_i) = P(X = x_i)$$



2. Cumulative Distribution Function

$$F_X(x_i) = P(X \leq x_i)$$



Intuitively, adding up $p_X(x_i)$ for all i is equal to $F_x(a)$.

$$F_X(a) = \sum_{\text{all } x_i \leq a} p_X(x_i)$$

Rules of Discrete Random Variables

- $0 \leq p_X(x_i) \leq 1$
- $\sum_{\text{all } x_i} p_X(x_i) = 1$
- $F_X(-\infty) = 0$
- $F_X(+\infty) = 1$
- $F_X(b) \geq F_X(a)$ if $b \geq a$

All of these rules are fairly intuitive. For example, the probability of any event must be between 0 and 1. Additionally, the sum of all events in a sample space must be 1.

2.2 Continuous Random Variables

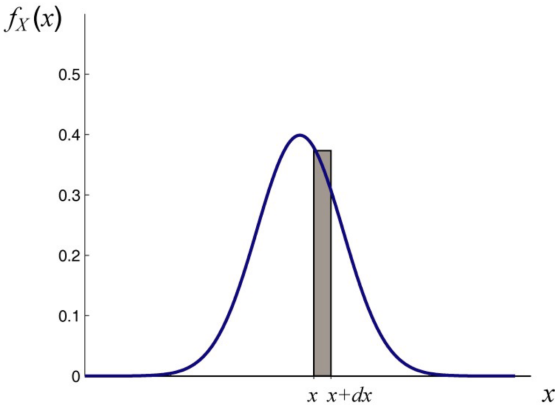
Random variables are said to be **continuous** if they can take on any real value. As a result, there are ∞ possible values for a random variable X. It follows that

$$P(X = x_i) = \frac{1}{\infty} = 0, \text{ for all } i$$

We can describe the distribution of continuous random variables in 2 ways.

1. Probability Density Function

$$f_X(x_i)dx = P(x_i < X < x_i + dx)$$

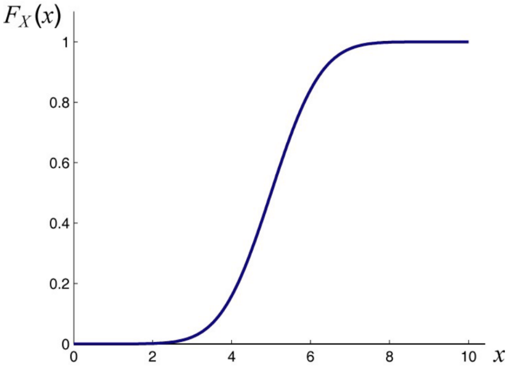


We know that occurences in different intervals are mutually exclusive, so it follows that

$$P(a < X \leq b) = \int_a^b f_X(x)dx$$

2. Cumulative Distribuion Function

$$F_X(x_i) = P(X \leq x_i)$$



Additionally,

$$F_X(x_i) = P(X \leq x_i) = \int_{-\infty}^{x_i} f_X(u)du$$

and it follows from the Fundamental Theorem of Calculus that

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

Rules of Continuous Random Variables

- $f_X(x) \geq 0$
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1 = S$
- $F_X(-\infty) = 0$
- $F_X(\infty) = 1$
- $F_X(b) \geq F_X(a)$ if $b \geq a$

Some of these rules hold in both the discrete and the continuous case.

2.3 Joint Random Variables

Sometime it is useful to consider the probabilistic relationship between two random variables. This is called the joint probability. Just as before, we can use 2 methods to describe the joint distribution of continuous random variables.

1. Joint Probability Density Function

   $$f_{X,Y}(x,y)dxdy = P(x < X \leq x + dx \cap y < Y \leq y + dy)$$

   Similarly to single PDF's, we can find the probability of X and Y within a certain region as follows.

   $$P(a < X \leq b \cap c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(u,v)dudv$$

   Two conditions must hold for joint PDF's

   (a) $f_{X,Y}(x,y) \geq 0$

   (b) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(u,v)dudv$

2. Joint Cumulative Distribution Function

   $$F_{X,Y}(x,y) = P(X \leq x \cap Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)dudv$$

   And just as in the single variable case, the Fundamental Theoreom of Calculus can be applied to obtain

   $$f_{X,Y}(x,y) = \frac{\delta^2}{\delta x \delta y} F_{X,Y}(x,y)$$

Marginal Distributions

Given the joint distribution of X and Y, we can obtain the distributions of X alone or Y alone. This is called the marginal distribution.

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

This follows from the fact that the the distribution of y from $-\infty$ to $+\infty$ is 1 and the intersection of any value with 1 is the original value.

The same style of manipulation can be performed in order to obtain the marginal CDF.

$$F_X(x) = F_X(x,\infty)$$

Conditional Probability Distributions

Conditional probability distributions are used to determine the probability of a variable given the value of another variable. Put another way, $P(X|Y)$.

For continuous random variables, the conditional PDF is

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

The conditional CDF is

$$F_{X|Y}(x|y) = P(X \leq x|Y = y) = \int_{-\infty}^{x} f_{X|Y}(u|y)du$$

Independence

Similar independence rules hold for random variables as in the case of events.

Specifically, X and Y are said to be independent if

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

The following are equivalent to the above statement

- $f_{Y|X}(y|x) = f_Y(y)$
- $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
- $F_{X|Y}(x|y) = F_X(x)$
- $F_{Y|X}(y|x) = F_Y(y)$
- $F_{X,Y}(x,y) = F_X(x)F_Y(y)$

We often assume independence in order to simplify calculations.

Joint Discrete Random Variables

The rules for joint distributions also apply to discrete random variables.

Specifically,

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

$$p_{X|Y} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

$$p_X(x) = \sum_{\text{all } y_i} p_{X,Y}(x,y_i)$$

$$p_{X|Y}(x|y) = p_X(X) \text{ given X and Y are independent}$$

3 | Expectations and Moments

Sometimes, it is convenient to use measures that describe the general features of a probability distribution, such as central location, breadth, and skewness. These features are called the **moments** of a random variable.

3.1 Expectations

Below we will explore several measures of central tendency as well has uncertainty.

Central Tendency

The most common of the central tendency measures is the **mean**, denoted as μ_x or $E[X]$. For discrete variables, it is calculated as follows

$$\mu_x = \sum_{\text{all } i} x_i p_X(x_i)$$

The continuous analog is

$$\mu_x = \int_{\text{all } x} x f_X(x) dx$$

Another central tendency measure is the **median**, denoted $x_{0.5}$. The median is defined as the as the value of X such that there is an equal probability that a random variable will fall below or above the value. It can be calculated using the CDF.

$$F_X(x_{0.5}) = 0.5$$

A final measure is the **mode**, denoted $\tilde{x}$. The value of the mode is the value which has the highest probability density, and there can be more than one. It can be calculated by finding the global maxima of the PDF function.

Generally the mode, median, and mean hold different values for an asymmetric probability distribution.

Uncertainty

The most common measure of uncertainty is the **variance**, denoted $var[X]$ or σ_X^2. The discrete case is below

$$\sigma_X^2 = \sum_{\text{all } i} (x_i - \mu_X)^2 p_X(x_i)$$

And the continuous case follows a similar idea

$$\sigma_X^2 = \int_{\text{all } x} (x - \mu_X)^2 f_X(x) dx$$

The square root of the variance is the **standard deviation**, denoted σ_X. This measure is generally more preferable than variance when reporting results due to the fact that it has the same units as the random variable (whereas the variance has the square of the original units).

Another option for measuring uncertainty is a unitless value called the **Coefficient of Variation**. This is denoted δ_X.

$$\delta_X = \frac{\sigma_X}{|\mu_X|}$$

This measure is useful for comparing random variables with different means, but does not work well when the mean is 0. It works best when the standard deviation is less than the mean.

Expectation Operator

Means and variances are special cases of the expectation operator. The expectation of g(X) is defined as

$$E[g(X)] = \sum_{\text{all } i} g(x_i) p_X(x_i)$$

$$E[g(X)] = \int_{\text{all } x} g(x) f_X(x) dx$$

By inspection, we can see that the mean is the case when $g(X) = X$ and the variance is the case when $g(X) = (X - \mu_X)^2$

Properties of Expectations

- $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$
- $E[cX] = cE[X]$
- $E[c] = c$
- $E[a + bx] = a + bE[x]$
- $var[cX] = c^2 var[X]$

These all follow from the linearity property of integrals and summations.

3.2 Moments

Given a function X^m, $E[X^m]$ is called the m$^{\text{th}}$ moment of X. The first few moments provide useful information about the distribution of X.

The function $(X - \mu_X)^m$ provides the central moments of X. When $m = 2$, this gives us the variance. When $m = 3$ and it is normalized by the standard deviation, we have a new measure called the **Coefficent of Skewness**.

$$\gamma = \frac{E[(X - \mu_X)^3]}{\sigma_X^3}$$

Positive values of γ indicate right skew while negative indicate left skew. The skew of data indicates the direction of a "long tail." The closer γ is to 0, the closer we are to a symmetric distribution.

When the 4$^{\text{th}}$ central moment is normalized by σ_X, we have the **Kurtosis Coefficient**, denoted κ. It is a measure of the length of the "tail" in the distribution.

Joint Moments

Joint moments can provide useful information about multiple random variables. The **covariance**, or the joint central moment of 2 random variables is computed

$$\sigma_{X,Y} = E[(X-\mu_X)(Y-\mu_y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_y)f_{X,Y}(x,y)dxdy$$

The covariance is a measure of the linear dependence between 2 variables. Using the linearity of the expectation operator we derive

$$\sigma_{X,Y} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

Properties of Covariance

- the covariance of X with itself is equal to the variance of X
- $|\sigma_{X,Y}| \leq \sqrt{\sigma_X^2\sigma_Y^2}$

Note: the covariance of X and Y is sometimes denoted Cov[X, Y].

If we normalize the covariance, we get the **correlation coefficent**.

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y}$$

The correlation coefficent is a dimensionless measure of linear dependence. When it equals 0, X and Y are said to be uncorrelated.

It is important to note that if X and Y are independent, they are uncorrelated but the reverse is not necessarily true. If X and Y are uncorrelated this does not imply independence. Two variables are independent if and only if P(A|B) = P(A).

3.3 Using Empirical Data

Probability Theory does not require real data. However, in our engineering applications of these tools we typically need to make use of observed data when creating our models. We need ways to organize and present our data so that it can be used effectively. These ideas fall under statistics rather than probability (recall that statistics is the field that treats past data while probability can be used to predict future events).

Empirical CDF

$$y_i = \frac{\text{No. of observations } \leq x}{\text{Total no. of observations}}$$

Numerical Summaries

Sample mean, median, and mode follow the typical definitions. That is the mean is the sum of all values divided by the number of values, the median is the middle value, and the mode is the most common value.

A measure of the variability of a sample is the sample variance, computed as follows:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample covariance and coefficent of correlation can be deduced in the same way.

4 | Probability Models

Up to this point, we have discussed general probability distributions. Now, we will discuss several specific probability distributions that are useful.

Bernoulli Distribution

The simplest scenario deals with a single trial that has a binary result, that is either a success or failure. For example, flipping a coin once.

Given a random variable X with x = {0, 1}, the PMF is

$$p_X(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Some more values of interest:

- Mean: $E[X] = p$
- Variance: $Var[X] = (1 - p)p$

Binomial Distribution of Bernoulli Sequence

We can repeat the above scenario for multiple trials. Give probability p of success for each trial and that each trial is independent, the number of successes for n trials is said to have a binomial distribution.

The PMF for n trials and probability p for each is

$$p_X(x) = \binom{n}{p} p^x (1 - p)^{n-x}$$

Some more values of interest:

- Mean: $E[X] = np$
- Variance: $Var[X] = np(1 - p)$
- Standard Deviation: $\sigma_X = \sqrt{np(1 - p)}$
- Coefficient of Variation: $\delta_X = \sqrt{\frac{1-p}{np}}$
- Skewness Coefficient: $\gamma_X = \sqrt{\frac{1-2p}{np(1-p)}}$

Geometric Distribution

Sometimes we want to know the probability of n failed trials before success. In other words, we want to know when the first sucess will occur. This is described by the geometric distribution.

This is given by

$$p_N(n) = p(1-p)^{n-1}$$

and the CDF

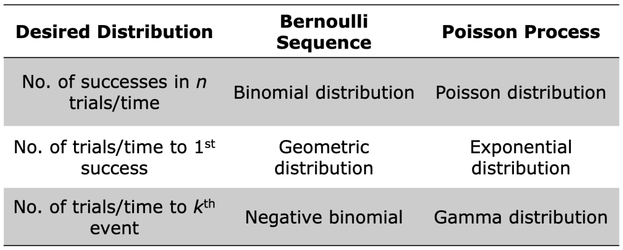
$$F_N(n) = 1-(1-p)^n$$

Some more values of interest:

- Mean: $\mu_x = \frac{1}{p}$
- Standard Deviation: $\sigma_N = \frac{\sqrt{1-p}}{p}$
- Coefficient of Variation: $\delta_N = \sqrt{1-p}$

Summary

- **Bernoulli Distribution**: a single trial where the outcome is either a success or a failure
  - the total number of successes, x, in n trials has a **binomial distribution**
  - the trial number, N, at which the first success occurs follows a **geometric distribution**
  - the trial number, W_k, at which the kth success occurs follows a **negative binomial distribution**
- **Poisson Distribution**: describes a random function w/ parameter lambda whose value at time t is number of incidents that have occured since $t=0$; the parameter v represents the expected number of events in the period t and λ is the rate of success per unit time
  - time to the first occurence of an event is represented by the **exponential distribution**
  - time to the kth occurence of an event is represented by the **gamma distribution**



Desired Distribution	Bernoulli Sequence	Poisson Process
No. of successes in *n* trials/time	Binomial distribution	Poisson distribution
No. of trials/time to 1st success	Geometric distribution	Exponential distribution
No. of trials/time to *k*th event	Negative binomial	Gamma distribution