Chapter 7; Common Models for an Uncertain World; Review of Probability

Steve Elston

12/9/2020

Introduction

Probability theory is the basis of statistics, machine learning and much of artificial intelligence. In this chapter we will review some basic principles and properties of probability that will be used throughout the rest of this book. However, this review is far from comprehensive, and you may wish to consult one of the many excellent introductory textbooks on probability theory for a comprehensive introduction.

There are a great many probability distributions which have been developed over more than two centuries. No book of reasonable length could discuss all of these in any depth. Some distributions are quite specialized and of limited general interest. In this chapter we will explore the basic properties of a few widely useful probability distributions. In subsequent chapters we will introduce several additional distributions as the need arises.

Early History of Probability

Jacob Bernoulli (1654, 1705) was a Swiss mathematician who pioneered many subjects in mathematics, including the mathematical theory of probability.

Figure 1: Jacob Bernoulli: Be happy he is not your statistics professor! Credit, Wikipedia commons

Bernoulli died before he could publish his book, *Artis conjectandi*. This book included a theory of probabilities from trials with discrete outcomes. His incomplete book was eventually published posthumously in 1713.

1

ARTIS CONJECTANDI
PARS PRIMA,
Complectens
Tractatum Hugenii de Ratiociniis
in Ludo Aleæ,
Cum Annotationibus
JACOBI BERNOULLJ.

CHRIST. HUGENII
ad
FRANC. SCHOOTENIUM
Præfatio.

CUm in editione elegantiſſimorum ingenii Tui monumentorum, quam præ manibus nunc habes, Vir Clariſſime, id inter cœtera Te ſpectare ſciam, ut varietate rerum, quarum tractationem inſtituiſti, oſtendas quàm latè ſe protendat divina Analytices ſcientia, facilè intelligo etiam illa plurimùm propoſito Tuo inſervire poſſe, quæ de Aleæ Ratiociniis

A con-

Figure 2: First probability textbook: How good is your Latin? Credit, Wikipedia commons

What Are Random Variables?

In statistics we often speak of **random variables**. But what does this mean exactly?
In simple terms, a random variable is any **mapping**, X, from from some outcome of a random event, ω, to a real number, $\mathbb{R}$. We can write this formally:

$$X(\omega) \rightarrow \mathbb{R}$$

This mapping can take many practical forms. For example, the mapping can be a count, or some kind of function which transforms ω to a real number, $\mathbb{R}$. This concept appears abstract at first glance, but is fundamental to the theory of probability. We will discuss a number of examples in this chapter and in subsequent chapters.

Discrete Probability Distributions and Random Variables for Counts

There are two classes of probability distributions; discrete and continuous. To get started, we will focus on discrete distributions.

Discrete distributions are used to model the probability of events with discrete outcomes or states. In simple terms, discrete distributions model probabilities for processes having a **countable possible outcomes**. Discrete distributions measure the probability of each of these outcomes, or possible counts. The probability of a discrete outcome is called the **Probability Mass Function**.

Axioms of probability

All probability distributions of any random variable must have certain properties. These properties are referred to as the **axioms of probability**. A discrete distribution is defined by its **probability mass function (PMF)**. A probability mass function must have non-zero values at each possible non-zero outcome. In this context the count of events is a random variable.

For discrete distributions, we can speak of a **set of events** within the **sample space** of all possible events. For discrete distributions the three axioms of probability then become:

1. Probability for any set of events, A, is greater than 0 and less than or equal to 1. This means that an event that can occur must have a non-zero probability, and that an event that must happen has a probability of 1. We can express this axiom mathematically as:

$$0 < P(A) \leq 1$$

2. The sum of the probability mass functions over the sample space must add to 1. In other words, the probability summed over all possible events must sum to 1. We can express this relationship mathematically as:

$$P(S) = \sum_{a_i \in A} P(a_i) = 1$$

Where the symbol $\forall$ is read **for all**.

3. If sets of events A and B are mutually exclusive, then the probability of either A and B is the probability of A plus the probability of B, since only one or the other can occur. This third axiom can be extended to any number of mutually exclusive sets. We express this axiom mathematically as:

$$P(A\ \cup B) = P(A) + P(B) if\ A \perp B$$

From these three axioms we can draw some useful conclusions. Events which cannot occur have probability 0. Further, events that must occur have probability 1. In any events, events must have a probability mass function between 0 and 1.

What do you expect: discrete distributions

In statistics it is often useful to know the value we should expect to find when we sample a random variable. We call this quantity the **expected value** or simply the **expectation**. If we have n samples, $\mathbf{X} = x_1, x_2, \ldots, x_n$, of a random variable with discrete probability mass function $p(x_i)$ the expected value of the sample is:

$$\mathrm{E}[\mathbf{X}] = \sum_{i-1}^{n} x_i\ p(x_i)$$

How can we interpret expectation? If you examine the relation above, you can see that expectation is a probability weighted sum of the sample of the random variable, $\mathbf{X}$. By the second axiom of probability presented above, the weights must sum to 1.0.

The above relationship is linear. Therefore, we can state some useful properties of expectation:
1. The expectation of the sum of two random variables, X and Y, is the sum of the expectations. This property is useful in computations, since we can compute expectations of two random variables independently. In mathematical terms this relationship is:

$$\mathrm{E}[\mathbf{X}, \mathbf{Y}] = \mathrm{E}[\mathbf{X}] + \mathrm{E}[\mathbf{Y}]$$

2. The expectation of an **affine transformation** of a random variable, X, is an **affine transformation** of the expectation. This properties is useful in computations since we can easily affine transform expectations without recomputing a known expectation. In mathematical terms this relationship is:

$$\mathrm{E}[\mathbf{a}\ \mathbf{X} + \mathbf{b}] = a\ \mathrm{E}[\mathbf{X}] + b$$

Bernoulli distributions

As a first example of a discrete distribution we will examine the properties of the **Bernoulli distribution**. Bernoulli distributions model the results of a **single trial** or **single realization** with a binary outcome. We call this single experiment a **Bernoulli trial**. A classic example of a Bernoulli trial is a single flip of a coin. The flip can only result in two possible outcomes, or end states, $\{heads, tails\}$.

For an event with a binary outcome, $\{0, 1\}$, or $\{failure, success\}$, with probability p of success, we can write the probability mass function for the Bernoulli distribution as:

$$
\begin{aligned}
P(x \mid p) &= \begin{cases} p\ if\ x = 1 \\ (1-p)\ if\ x = 0 \end{cases} && (1)\\
or & && (2)\\
P(x \mid p) &= p^x (1-p)^{(1-x)}\ x \in 0, 1 && (3)
\end{aligned}
$$

Some other basic properties of the Bernoulli distribution are:

$$
\begin{aligned}
Mean &= p && (4)\\
Variance &= p(1-p) && (5)
\end{aligned}
$$

Binomial distribution

As discussed, the Bernoulli distribution only applies to a single experiment or trial. This idealized situation is rarely seen in the real-world. How do we deal with situations where we must model the number of successful outcome in N trials? In these cases we use the **Binomial distribution**.

The Binomial distribution is widely used in statistics, classification in machine learning, and reasoning problems in artificial intelligence. Any case where where there are a series of experiments with two possible outcomes the Binomial distribution is useful. The mostly widely used model with Binomially distributed outcomes is logistic regression.

You can think of the Binomial distribution and the product of multiple Bernoulli trials. For example, if we perform N Bernoulli trials with outcomes, $(\{success, fail\})$, on a sample of trials (with replacement). The probability of k successes in N Bernoulli trials with probability of positive outcome p is then written as:

$$P(k \mid N, p) = \binom{N}{k} p^k (1-p)^{(N-k)}$$

The product of Bernoulli trials is normalized by the **Binomial coefficient**. This normalization accounts for all possible combinations of outcomes from the trials, and ensures the distribution is in the range $0 \leq P(k \mid N, p) \leq 1$. The value of the Binomial coefficient is the number of ways k items can be selected from N possibilities.

How do we compute the expected value of successes in a sequence of N trials. The probability of success on each trail is p, a constant. If x represents a single trial, the expected number of successes is:

$$
\begin{aligned}
\mathrm{E}_N &= \sum_{i-1}^{N} x_i \; p(x_i) && (6)\\
&= \sum_{i-1}^{N} p \; x_i && (7)\\
&= \sum_{i-1}^{n} p && (8)\\
&= p \; N && (9)
\end{aligned}
$$

The mean and variance are just the Bernoulli mean and variance multiplied by the number of trials:

$$
\begin{aligned}
Mean &= Np && (10)\\
Variance &= Np(1-p) && (11)
\end{aligned}
$$

> **Exercise 7-1:** You will compute and compare several sets of realizations of a Binomial distributions, with probability of success $p = 0.75$, and $N = \{5, 25, 75, 100\}$. For each value of N, compute 1000 independent samples. You may use the numpy.random.binomial function to compute these realizations. Then do the following:
> 1. Compute and print the theoretical and sample means and variances for each case of 1000 realizations. Are the sample means and variances close to the theoretical values? Does the correspondance between the sample values and theortical values improve with sample size, and what does this tell you about working with small and large samples?
> 2. Create and execute the code to plot density histograms of the four sets of realizations of 1000 on a 2x2 grid of plots. 3. On the histogram plots supperimpose a plot of the density on the Normal distribution with the theoretical mean and variance, using 100 points.
> 4. Do these distributions appear as you expect? Do the histograms appear to converge to the 'bell-shapped curve' of the Normal distribution?

Poisson Distribution

Modeling the counts of events occurring within a period of time is a common problem. The Poisson distribution models the occurrence of events in a fixed interval of time. We say that the Poisson distribution models the probability, P, of x **arrivals** within the time period.

A Poisson process is an example of a **point process**. A temporal point process models the probability of a number of events (points) occurring in a time period. In point process terminology, the average number of arrivals of the Poisson process is referred to as the **intensity of the process**.

Point process models are widely used. Examples include reliability analysis in engineering, survival of patients receiving medical treatments, customer loyalty and epidemiology. All of these problems involve modeling the intensity of some process.

In mathematical terms we write the Poisson distribution in terms of the average arrival rate, λ as:

$$P(x \mid \lambda) = \frac{\lambda^x}{x!} \exp^{-\lambda}$$

The mean and variance of the Poisson distribution are both equal to the parameter λ, or:

$$Mean = \lambda \tag{12}$$
$$Variance = \lambda \tag{13}$$

Exercise 7-2: Create and execute code to compute 1000 realizations of Poisson distributions with average arrival rates, $\lambda = \{1, 5, 25, 100\}$. You can use the numpy.random.poisson function to compute these realizations.
1. Compute and print the sample mean and variance for each set of realization along with the theoretical values. Are the sample means and variances close to the theoretical values? Does the correspondance between the sample values and theortical values improve with sample size, and what does this tell you about working with small and large samples?
2. Create and execute the code to plot the density histograms of the four sets of realizations of 1000 on a 2x2 grid of plots. 3. On the histogram plots supperimpose a plot of the density on the Normal distribution with the theoretical mean and variance, using 100 points.
4. Do these distributions appear as you expect? Do the histograms appear to converge to the 'bell-shapped curve' of the Normal distribution?

Continuous Distributions

Continuous distributions are used to model continuous valued random variables. Physical measurements, such as weight, length and temperature, are examples of variables with continuous variables. In practice, we treat random variables with a great number of discrete values as continuous. Examples of the later include, prices of items or assets and units of industrial production.

Continuous distributions have an infinite number of possible outcomes. Therefore, probability must be measured for some range of values or **finite interval**. We therefore call the distribution function of continuous random variables the **Probability Density Function (PDF)**. This is in contrast to the probability mass function for discrete distributions.

When working with continuous distributions it is important to keep the nature of the probability density function in mind. First, The probability of an interval, $\{X_1, X_2\}$, of a random variable equals the **area** under density curve over that interval:

$$P(X_1, X_2) = \int_{X_1}^{X_2} P(x) dx$$

As a consequence of the foregoing, the density function at a single value has infinitesimal mass. Therefore, the probability of any single, exact value is 0:

$$\int_{X_1}^{X_1} P(x) dx = 0$$

Axioms of probability for continuous distribtions

For continuous distributions we can state the three **axioms of probability** as:

1. Probability on the an interval, $\{X_1, X_2\}$, for a PDF, $P(x)$, must be bounded by 0 and 1:

$$0 \leq \int_{X_1}^{X_2} P(x) dx \ \leq 1$$

2. The area under the entire PDF must be equal to 1, integrated over the range of possible values:

$$\int_{-\infty}^{\infty} P(x)dx = 1$$

It is important to keep in mind, that for some distributions the PDF is not defined on $x < 0$.

3. If events A and B are mutually exclusive, then the probability of either A or B is the probability of A plus the probability of B:

$$P(A\ \cup B) = P(A) + P(B) if\ A \perp B$$

What do you expect: continuous distributions

We have already discussed **expected value** for discrete distributions. The same concept applies to continuous distributions. For the continuous distribution case we use a PDF. If we have samples, $\mathbf{X}$, of a continuous random variable with probability density function $p(x)$, the expected value over the interval, $\{a, b\}$, is:

$$\mathrm{E}[\mathbf{X}] = \int_a^b x\ p(x)\ dx$$

The interpretation of the expectation of a continuous distribution is similar to the discrete case. The values x are weighted by the PDF, $p(x)$. By the second axiom of probability presented above, PDF must equal 1.0 integrated over the entire range of x.

The two properties of expectations for the sum of random variables and the affine transformation of a random variable, discussed discrete random variables, also apply to continuous random variables.

Uniform distribution

A Uniform distribution has a flat PDF between limits $\{a, b\}$ and 0 outside that interval. The Uniform distribution is used in a number of important applications. Uniform distributions are fundamental to random sampling of data and in simulation. Further, transformations of the Uniform distribution are typically used to generate realizations of other distributions in computational statistics.

We can write the probability of the the Uniform distribution as:

$$P(x\ |\{a, b\}) = \begin{cases} \frac{1}{(b-a)}\ if\ a \leq x \leq b \\ else\ 0 \end{cases}$$

The Uniform distribution has the following properties:

$$Mean = \frac{(a+b)}{2} \tag{14}$$

$$Variance = \frac{1}{2}(b-a)^2 \tag{15}$$

The expectation of a uniform distribution on the interval $\{a, b\}$ is easy to work out:

$$
\begin{aligned}
\mathrm{E}_{a,b}(\mathbf{X}) &= \int_a^b x\ p(x)\ dx && (16)\\
&= \int_a^b x\ dx && (17)\\
&= \frac{x^2}{2}\ |_a^b && (18)\\
&= \frac{a+b}{2} && (19)
\end{aligned}
$$

Which is just the mean.

> **Exercise 7-3:** Create and execute code to compute and plot histograms with kernel density estimates of the uniform distribution on the interval $\{0, 1\}$ for $\{100, 1000, 10000, 100000\}$ realizations on a 4x4 grid. You can compute the realizations using the numpy.random.uniform function. How close to the ideal Uniform distribution are these different realizations?

Normal distribution

The **Normal distribution** or **Gaussian distribution** is one of the most widely used probability distributions. For most any case which are the product of a large number of processes or where large numbers of samples are available, random variables converge to a Normal distribution. The **central limit theorem**, which we address in Chapter XXXX, is a key example.

Many physical processes produce measurement values which are reasonably well modeled by a Normal distribution or the Log-Normal distribution. Further, the Normal distribution has tractable mathematical properties. In statistics and machine learning the response variable in linear regression is modeling using a Normal distribution. But, the Normal distribution is important in a much wider range of applications. For example, Normal distributions are used for navigation problems like GPS location, telecommunications signal analysis. We will explore additional properties and applications of the Normal distribution in subsequent chapters.

For a univariate Normal distribution we can write the density function as:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

The parameters can be interpreted as:

$$
\begin{aligned}
\mu &= location\ parameter = mean && (20)\\
\sigma &= scale = standard\ deviation && (21)\\
\sigma^2 &= Variance && (22)
\end{aligned}
$$

In other words, the location parameter determines the center of the distribution. The scale parameter determines spread or dispersion of the distribution. The special case of $\mu = 0$ and $scale = \sigma = 1$ is known as the **standard Normal**.

What about the expected value? For the case of a **standard Normal**, $\mathcal{N}(0, 1)$,We can work this out as follows:

$$
\begin{aligned}
\mathrm{E}[\mathcal{N}(0,1)] &= \int_{-\infty}^{\infty} x \ \frac{1}{\sqrt{2\pi}} \exp{\frac{-x^2}{2}} \ dx && (23)\\
&= \frac{1}{\sqrt{2\pi}} \exp{\frac{-x^2}{2}} |_{-\infty}^{\infty} && (24)\\
&= 0 && (25)
\end{aligned}
$$

For a Normal distribution with mean μ, $\mathcal{N}(\mu, 1)$, we can find the expectation is μ by applying the affine transformation property:

$$\mathrm{E}[\mathcal{N}(\mu, 1)] = \mathrm{E}[\mathcal{N}(0, 1) + \mu] = \mathrm{E}[\mathcal{N}(0, 1)] + \mu = 0 + \mu = \mu$$

Exercise 7-4: for the four sets of parameters shown in the table below, compute 10000 realizaitons of the distribution and plot the density function on the interval $-3 \leq x \leq 9$. You can use the numpy.random.normal function. Make sure you use distinct lines for these plots. How does the location and scale change with the two parameters?

μ	σ
0	1
5	1
0	0.1
4	4

Exercise 7-5: The density function of the Normal distribution exhibits the famous **'bell-shaped curve'**. But, how does the shape become apparent as the number of samples increases? Execute the code in the cell below to find out. Compute and plot histograms and density estiamtes for $[100, 1000, 10000, 100000]$ realizations. How does the distribution change with the number of realizations. Does the distribution approach the ideal 'bell shaped curve' as the number of samples increases?

Log-Normal distribution

The Normal distribution is defined for continuous random variables in the range $-\infty \leq x \leq \infty$. However, many continuous random variables are only defined in a range $0 < x \leq \infty$. Examples include, price, weight, length, and volume. In many of these cases the **Log-Normal** distribution is a good choice for the data generating process.

The Log-Normal distribution is based on a log-transformation of the random variable. The probability density function is:

$$P(x) = \frac{1}{x} \frac{1}{\sigma \sqrt{2\pi}} \exp{\frac{-(log(x) - \mu)^2}{2\sigma^2}}$$

Exercise 7-6: You will now make two side-by-side plots to compare a log-distributed variable and the same variable log-trnsformed. 1. Compute and plot a histogram and the kernel density estimate of 100000 realizations of a standard Log-Normal distribution using the scipy.stats.lognorm.pdf function.
2. Now log-transform the realizations of the log-Normal random variable and plot a histogram and the kernel density estimate. 3. In a seperate plot area, display a q-q Normal plot of the

> log-transformed random variable.
> 4. Answer these questions. Is the long right tail as expected in a log-distributed random variable? Does the log-transformed variable PDF plot resemble the Normal distribution?

Student t-distribution

The **Student t-distribution**, often just referred to as simply the t-distribution, is of importance in statistics since the difference of means of two Normally distributed random variables is t-distributed. This property makes the t-distribution important in hypothesis testing, which is discussed in Chapter XXX.

The t-distribution is defined in a somewhat different way from the other distributions we have looked at. It has one parameter, the **degrees of freedom**, denoted as ν. The derivation of the density function for the t-distribution is a bit complicated and leads to the following, rather complex looking result:

$$P(x \mid \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} where \Gamma(x) = Gamma\ function$$

> **Exercise 7-7:** Despite the complex density formula, you can gain a fair feel for the behavior of the t-distribution by plotting it for several values of the degrees of freedom, ν, and comparing it to the Normal distribution. Create and execute code to plot the density function for the t-distribution for degrees of freedom, $\nu = \{1, 2, 4, 8, \infty\}$, along with the density of the standard Normal distribution for 100 points on the interval $\{-4, 4\}$. You can use the scipy.stats.pdf function. Make sure you use distinct lines and include a legend for each. Answer these questions:
> 1. How do the shape (body and tails) of the t-distribution change with ν?
> 2. How would you describe the convergence of the t-distribution to the Normal with changing ν?

The Gamma and χ^2 distributions

The **Gamma family of distributions** includes several members which are important in statistics. As with the Normal distribution, the Gamma distributions are a two-parameter exponential family. An important property of the Gamma family is that the PDF is defined in the range $0 \leq x \leq \infty$. As a result, members of the Gamma family are used in many problems, ranging from measurements of physical systems to hypothesis testing.

The Gamma family can be in several ways. In this discussion we will use a with a shape parameter, ν and a scale parameter σ. Alternatively, one could use an inverse scale parameter, $\beta = 1/\sigma$. Using our chosen parameterization, we write the PDF of the Gamma distribution:

$$Gam(\nu, \sigma) = \frac{x^{\nu-1}\ e^{-x/\sigma}}{\sigma^{\nu}\ \Gamma(\nu)} where x \geq 0,\ \nu > 0,\ \sigma > 0 and \Gamma(\nu) = Gamma\ function$$

Two useful special cases of the Gamma distribution are:

1. $Gam(1, 1/\lambda)$ is the **exponential distribution** with decay constant λ, and PDF:

$$exp(\lambda) = \lambda e^{-\lambda x}$$

2. $Gam(\nu/2, 2) = \chi^2_\nu$ is the **Chi-squared distribution** with ν degrees of freedom. The χ^2_ν distribution has many uses in statistics. One important use is the distribution of estimates of the variance of the Normal distribution. Using the relationship to the Gamma distribution, we can then write the PDF of the χ^2_ν distribution:

$$\chi^2_\nu = \frac{x^{\nu/2-1}\ e^{-x}}{\sigma^{\nu/2}\ \Gamma(\nu/2)} for\ \nu\ degrees\ of\ freedom$$

> **Exercise 7-8:** To gain a feel for the χ^2_ν distribution you will plot it for 100 points on the interval $\{0 \leq x \leq 9\}$ for $\nu = \{1, 2, 3, 6, 9\}$ degrees of freedom. Make sure the lines for each degree of freedom are unique, and include a legend.

Multivariate Distributions

In all of the foregoing, we have only addressed distributions of a single real-valued random variable on $\mathbb{R}$. In a great many practical applications the random variable of interest is an n-dimensional vector in $\mathbb{R}^n$. As a result, range of applications of multi-variate distributions is immense. In these situations, we model the random variable using a **multivariate distribution**.

For now, we will only look briefly at one example, the n-dimensional **multivariate Normal distribution**. This distribution has two multi-valued parameters; an n-dimensional vector of locations, $\vec{\mu}$ and an n x n dimensional **covariance matrix**, $\mathbf{\Sigma}$; $\mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$. For k vector-valued samples, $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k\}$, we can can express the values of the covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix}
\sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,k} \\
\sigma_{2,1} & \sigma_{2,2} & \dots & \sigma_{2,k} \\
\vdots & \vdots & \vdots & \vdots \\
\sigma_{k,1} & \sigma_{k,2} & \dots & \sigma_{k,k}
\end{bmatrix}$$

For a Normally distributed k-dimensional multivariate random variable the terms, $\sigma_{i,j}$, are computed from the sample, $\mathbf{X}$, using the following relationship:

$$\begin{aligned}
\sigma_{i,j} &= \mathrm{E}\big[(\vec{x}_i - \mathrm{E}[\vec{x}_i]) \cdot (\vec{x}_j - \mathrm{E}[\vec{x}_j])\big] && (26) \\
&= \mathrm{E}\big[(\vec{x}_i - \bar{x}_i) \cdot (\vec{x}_j - \bar{x}_j)\big] && (27) \\
&= \frac{1}{k}(\vec{x}_i - \bar{x}_i) \cdot (\vec{x}_j - \bar{x}_j) && (28)
\end{aligned}$$

Where $\cdot$ is the inner product operator and $\bar{x}_i$ is the mean of $\vec{x}_i$.

Putting the above together, we can express the PDF of the multivariate Normal distribution as:

$$f(\vec{\mathbf{x}}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} exp\big(\frac{1}{2}(\vec{\mathbf{x}} - \vec{\mu})^T \mathbf{\Sigma} (\vec{\mathbf{x}} - \vec{\mu})\big)$$

Where $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix.

Notice that along the diagonal the values are the n variances of each dimension, $\sigma_{i,i}$. The off-diagonal terms describe the **dependency** between the n dimensions of the distribution.

> **Exercise 7-9:** In some cases, there is little or no dependency between the n dimensions of the distribution. How does this **statistical independence** change the on-diagonal and off-diagonal elements of the covariance matrix? Does it appear that this simplification leaves a model with n univariate Normal random variables?

Distributions for Multiple Outomes; the Categorical and Multinomial Distribution

We have already examined the Bernoulli and Binomial distributions. These distributions model cases where there are two possible outcomes, $\{0, 1\}$, or $\{success, failure\}$. But, many real-world cases have many possible

outcomes. Just of few of the many possible examples include:
1. The side which comes up when rolling a 6-sided dice.
2. A customer needing a hammer can choose between many possible types and manufacturers.
3. A bird seen in an image could be one of thousands of possible species.
4. Given a vague set of symptoms, a patient may have any one of several possible diseases.

In these cases we need a probability distribution for multiple outcomes. This is where the **Categorical distribution** comes into play. The Categorical Distribution is the multiple-outcome extension of the Bernoulli distribution, and is sometimes call the **Multinoulli distribution**.

The Categorical distribution

Let's say that we have a sample space of k possible outcomes, $\mathcal{X} = (e_1, e_2, \ldots, e_k)$. For each trial, there can only be one outcome. For outcome i we can encode the results as:

$$\mathbf{e_i} = (0, 0, \ldots, 1, \ldots, 0)$$

Where only value e_i has a value of 1. This representation is known as **one hot encoding** since only one value is nonzero at a time.

For a single trial the probabilities of the k possible outcomes can be expressed:

$$\Pi = (\pi_1, \pi_2, \ldots, \pi_k) \tag{29}$$
$$with \sum_i \pi_i = 1 \tag{30}$$

And consequently, we can write the simple probability mass function as:

$$f(x_i|\Pi) = \pi_i$$

For a series n of trials we can estimate each of the probabilities of the possible outcomes, $(\pi_1, \pi_2, \ldots, \pi_k)$:

$$\pi_i = \frac{\#\ e_i}{n}$$

Where $\#\ e_i$ is the count of outcome e_i.

For the case of $k = 3$ you can visualize the possible outcomes of a single Categorical trial. Each discrete outcome must fall at one of the corners of a **simplex**, as shown in Figure 3. The probabilities of of each outcome is (π_1, π_2, π_3).

Relationship to the Bernoulli distribution

How is the Categorical distribution related to the Bernoulli distribution? Recall that the Bernoulli distribution has probabilities of failure and success, $\{f, s\}$:

$$\Pi = (\pi_f, \pi_s) \tag{31}$$
$$= (1 - p, p) \tag{32}$$
$$where,\ p = probability\ of\ success \tag{33}$$

From the foregoing, you can see that the Categorical distribution is identical to the Bernoulli distribution for $k = 2$.

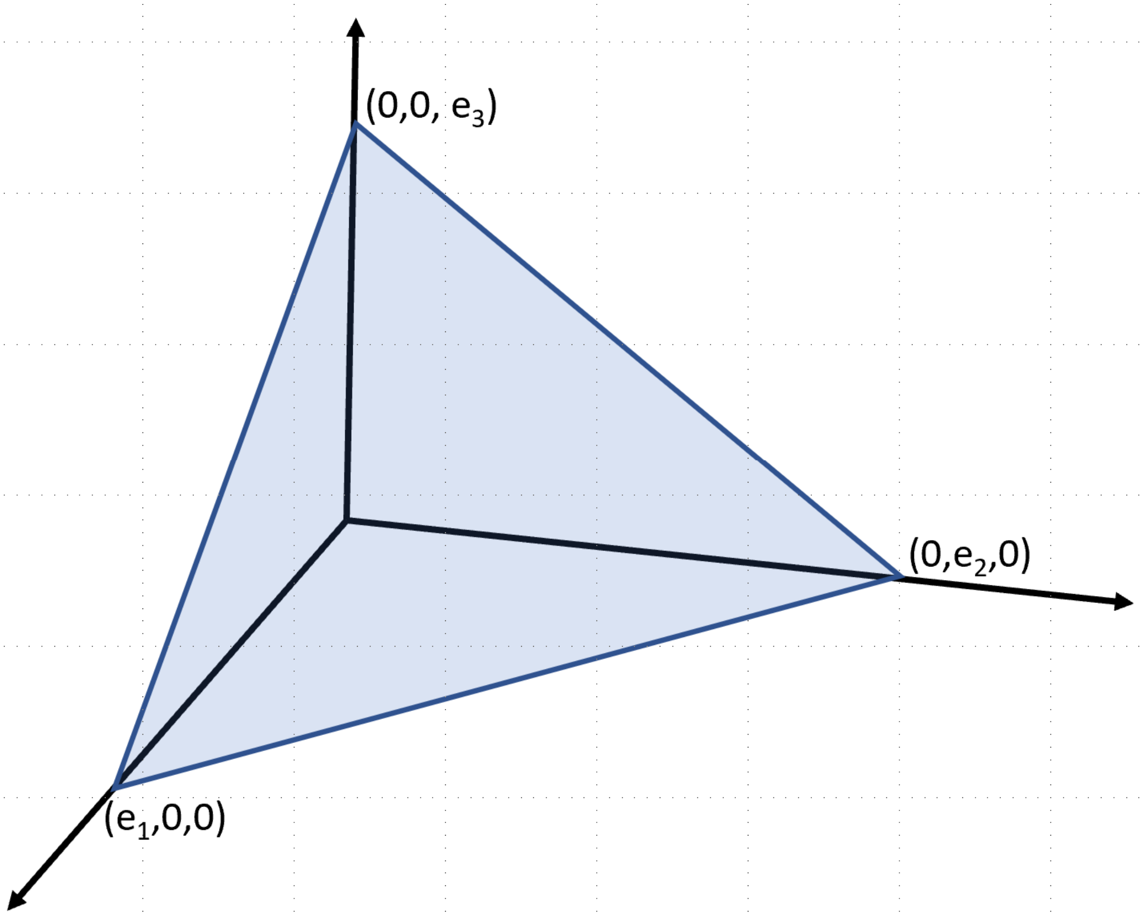

Figure 3: Simplex for $Mult_3$

The Multinomial distribution

How can we model the probability of a sequence of outcomes, $(x_1, x_2, \ldots, x_k)$, where $x_i = \#\ e_i$, the count of events, e_i? Generating such a sequence is equivalent to performing n single categorical trials. The probability of generating this squence of counts is:

$$p(x_1, x_2, \ldots, x_k; n, \pi_1, \pi_2, \ldots, \pi_k) = \frac{n!}{x_1!x_2!\ldots, x_k!}\pi_1^{x_1}, \pi_2^{x_2}, \ldots, \pi_k^{x_k}$$

This is formatible equation! Further, computing such a quantity for more than simple cases can be difficult, even prohibitive. We can write the forgoing in a somewhat more tractable form using gamma functions:

$$p(x_1, x_2, \ldots, x_k; n, \pi_1, \pi_2, \ldots, \pi_k) = \frac{\prod_i \left(\sum_i \Gamma(x_i + 1)\right)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} \pi_i^{x_i}$$

This is still a difficult relationship to work with. Fortunately, in many practical cases we can work with the far simpler categorical distribution.

> **Exercise 7-10:** All possible outcomes of 4 trails of a Categorical distribution with $k = 3$ are shown on the simplex in Figure 4. Assume a series of 4 trials is performed. Consider three possible sets of outcomes shown on the Figure, A, B, and C. For each of these compute the probability vector Π.

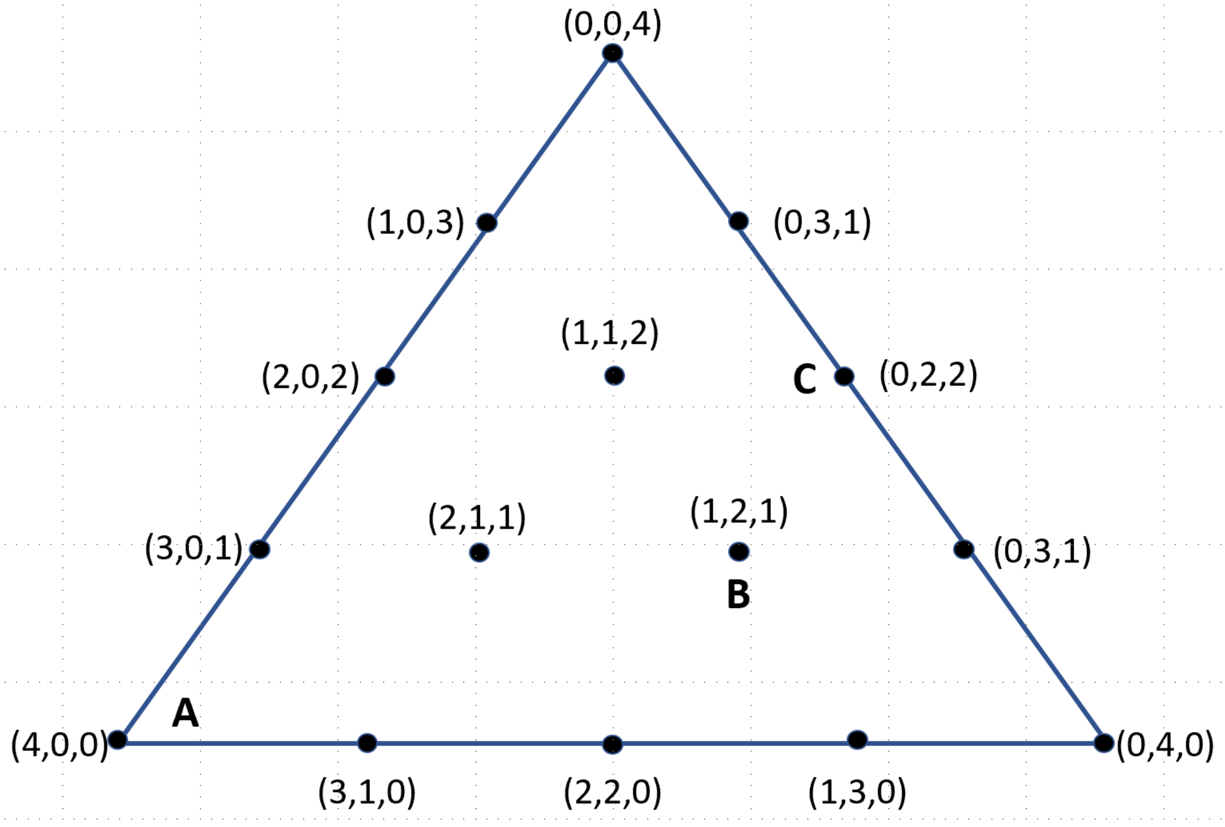

Figure 4: Sample space, (x_1, x_2, x_3), of $Mult_3$ with 4 samples

Odds

There is one more topic we will address to conclude our review of probability theory. **Odds** are the ratio of the number of ways an event occurs to the number of ways it does not occur. Sometimes we say that **odds** are the count of events in favor of an event vs. the count against the event. In subsequent chapters we will use odds (or more specifically log-odds) for evaluating and comparing models. For now, we will just introduce the concept.

If you flip a fair coin the odds of getting heads are $1:1$ (1 in 1). Since there are only two possible equally likely outcomes for the coin flip we say the odds of the outcome are even. As another example, if you roll a single fair die your odds of rolling a 6 are $1:5$ (1 in 5), or 0.2.

It is natural to ask, what is the relationship between odds and probability of an event? We can work this out for some event with count A in a set of all outcomes with count S, and where the count of negative outcomes $B = S - A$, as follows:

$$P(A) = \frac{A}{S} = \frac{A}{A + (S - A)} = \frac{A}{A + B} = \frac{count\ in\ favor}{count\ in\ favor\ + count\ not\ in\ favor} which\, implies odds = A : (S{-}A)$$

Let's say that for the fair coin flip, the odds are $1:1$. So we can compute the probability of heads as:

$$P(H) = \frac{1}{1+1} = \frac{1}{2}$$

As an example of the use of odds in statistics, consider the **odds ratio**, $\frac{p}{1-p}$, which is used to predict the response variable in logistic regression. This important transformation will be discussed in Chapter XXX of this book.

> **Exercise 7-11:** Answer the following questions:
> 1. We have said that the odds of rolling a 6 when throwing a single fair die are 1:5. Using these odds as the basis of your calculation, what is the probability of rolling a 6?
> 2. What are the odds and probability of rolling a total of 7 when throwing two fair dice?

Bibliography