1 Probability Review

Probability is a critical tool for modern data analysis. It arises in dealing with uncertainty, in randomized algorithms, and in Bayesian analysis. To understand any of these concepts correctly, it is paramount to have a solid and rigorous statistical foundation. Here we review some key definitions.

1.1 Sample Spaces

We define probability through set theory, starting with a *sample space* Ω. This represents the space of all things that might happen in the setting we consider. One such potential outcome $\omega \in \Omega$ is a *sample outcome*, it is an element of the space Ω. We are usually interested in an *event* that is a subset $A \subseteq \Omega$ of the sample space.

Example: Discrete Sample Space

Consider rolling a single fair, 6-sided die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. One roll may produce an outcome $\omega = 3$, rolling a 3. An event might be $A = \{1, 3, 5\}$, any odd number.
The probability of rolling an odd number is then $\textbf{Pr}(A) = |\{1, 3, 5\}|/|\{1, 2, 3, 4, 5, 6\}| = 1/2$.

A *random variable* $X : \Omega \to S$ is a **function** from the sample space Ω to a domain S. Many times $S \subseteq \mathbb{R}$, where $\mathbb{R}$ is the space of real numbers.

Example: Random Variable

Consider flipping a fair coin with $\Omega = \{H, T\}$. If I get a head H, then I get 1 point, and if I get a T, then I get 4 points. This describes the random variable X, defined $X(H) = 1$ and $X(T) = 4$.

The *probability* of an event $\textbf{Pr}(A)$ satisfies the following properties:

- $0 \leq \textbf{Pr}(A) \leq 1$ for any A,
- $\textbf{Pr}(\Omega) = 1$, and
- The probability of the union of disjoint events is equivalent to the sum of their individual probabilities. Formally, for any sequence $A_1, A_2, \ldots$ where for all $i \neq j$ that $A_i \cap A_j = \emptyset$, then

$$\textbf{Pr}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \textbf{Pr}(A_i).$$

Sample spaces Ω can also be continuous, representing some quantity like water, time, or land mass which does not have discrete quantities. All of the above definitions hold for this setting.

Example: Continuous Sample Space

Assume you are riding a Swiss train that is always on time, but its departure is only specified to the minute (specifically, 1:37 pm). The true departure is then in the state space $\Omega = [1{:}37{:}00, 1{:}38{:}00)$.
A continuous event may be $A = [1{:}37{:}00 - 1{:}37{:}40)$, the first 40 seconds of that minute.
Perhaps the train operators are risk averse, so $\textbf{Pr}(A) = 0.80$. That indicates that 0.8 fraction of trains depart in the first $2/3$ of that minute (less than the 0.666 expected from a uniform distribution).

1.2 Conditional Probability and Independence

Now consider two events A and B. The *conditional probability* of A given B is written $\mathbf{Pr}(A \mid B)$, and can be interpreted as the probability of A, restricted to the setting where we know B is true. It is defined in simpler terms as $\mathbf{Pr}(A \mid B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}$, that is the probability A and B are both true, divided by (normalized by) the probability B is true.

Two **events** A and B are *independent* of each other if and only if

$$\mathbf{Pr}(A \mid B) = \mathbf{Pr}(A).$$

Equivalently they are independent if and only if $\mathbf{Pr}(B \mid A) = \mathbf{Pr}(B)$ or $\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\mathbf{Pr}(B)$. By algebraic manipulation, it is not hard to see these are all equivalent properties. This implies that knowledge about B has no effect on the probability of A (and vice versa from A to B).

Example: Conditional Probability

Consider the two random variables. T is 1 if a test for cancer is positive, and 0 otherwise. Variable C is 1 if a patient has cancer, and 0 otherwise. The joint probability of the events is captured in the following table:

	$C = 1$	$C = 0$
$T = 1$	0.1	0.02
$T = 0$	0.05	0.83

Note that the sum of all cells (the joint sample space Ω) is 1. The conditional probability of having cancer, given a positive test is $\mathbf{Pr}(C = 1 \mid T = 1) = \frac{0.1}{0.1+0.02} = 0.8333$. The probability of cancer (ignoring the test) is $\mathbf{Pr}(C = 1) = 0.1 + 0.05 = 0.15$. Since $\mathbf{Pr}(C = 1 \mid T = 1) \neq \mathbf{Pr}(C = 1)$, then events $T = 1$ and $C = 1$ are not independent.

Two **random variables** X and Y are *independent* if and only if, for *all* possible events $A \subseteq \Omega_X$ and $B \subseteq \Omega_Y$ that A and B are independent: $\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\mathbf{Pr}(B)$.

1.3 Density Functions

Discrete random variables can often be defined through tables (as in the above cancer example). Or we can define a function $f_X(k)$ as the probability that random variable X is equal to k. For continuous random variables we need to be more careful; we use calculus. We will next develop probability density functions and cumulative density functions for continuous random variables; the same constructions are sometimes useful for discrete random variables as well, which basically just replace a integral with a sum.

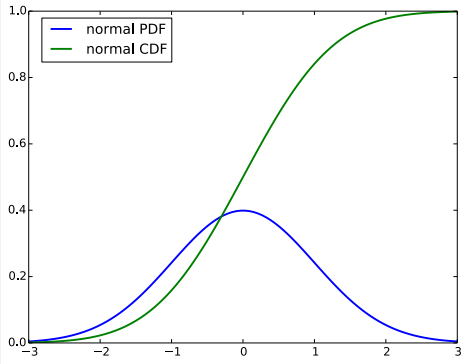
We consider a continuous sample space Ω, and a random variable X defined on that sample space. The probability density function of a random variable X is written f_X. It is defined with respect to any event A so that $\mathbf{Pr}(X \in A) = \int_{\omega \in A} f_X(\omega) \mathrm{d}\omega$. The value $f_X(\omega)$ *is not equal to* $\mathbf{Pr}(X = \omega)$ in general, since for continuous functions $\mathbf{Pr}(X = \omega) = 0$ for any single value $\omega \in \Omega$. Yet, we can interpret f_X as a *likelihood* function; its value has no units, but they can be compared and larger ones are more likely.

Next we will defined the *cumulative density function* $F_X(t)$; it is the probability that X takes on a value of t or smaller. Here it is typical to have $\Omega = \mathbb{R}$, the set of real numbers. Now define $F_X(t) = \int_{\omega=-\infty}^{t} f_X(\omega) \mathrm{d}\omega$.

We can also define a pdf in terms of a cdf as $f_X(\omega) = \frac{\mathrm{d}F_X(\omega)}{\mathrm{d}\omega}$.

Example: Normal Random Variable

A *normal* random variable X is a very common distribution to model noise. It has domain $\Omega = \mathbb{R}$. Its pdf is defined $f_X(\omega) = \frac{1}{\sqrt{2\pi}} \exp(-\omega^2/2) = \frac{1}{\sqrt{2\pi}} e^{-\omega^2/2}$, and its cdf has no closed form solution. We have plotted the cdf and pdf in the range $[-3, 3]$ where most of the mass lies:




```
import matplotlib as mpl
mpl.use('PDF')
import matplotlib.pyplot as plt
from scipy.stats import norm
import numpy as np
import math

mu = 0
variance = 1
sigma = math.sqrt(variance)
x = np.linspace(-3, 3, 201)

plt.plot(x, norm.pdf((x-mu)/sigma),linewidth=2.0, label='normal␣PDF')
plt.plot(x, norm.cdf((x-mu)/sigma),linewidth=2.0, label='normal␣CDF')
plt.legend(bbox_to_anchor=(.35,1))

plt.savefig('Gaussian.pdf', bbox_inches='tight')
```

1.4 Expected Value

The expected value of a random variable X in a domain Ω is a very important constant, basically a weighted average of Ω, weighted by the range of X. For a discrete random variable X it is defined

$$\mathbf{E}[X] = \sum_{\omega \in \Omega} \omega \cdot \mathbf{Pr}[X = \omega].$$

For a continuous random variable X it is defined

$$\mathbf{E}[X] = \int_{\omega \in \Omega} \omega f_X(\omega) \mathrm{d}\omega.$$

Linearity of Expectation: An important property of expectation is that it is a linear operation. That means for two random variables X and Y we have $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$. For a scalar value α, we also $\mathbf{E}[\alpha X] = \alpha \mathbf{E}[X]$.

Example: Expectation

Let H be the random variable of the height of a man in meters without shoes. Let the pdf f_H of H be a normal distribution with expected value $\mu = 1.755$m and with standard deviation 0.1m. Let S be the random variable of the height added by wearing a pair of shoes in centimeters (1 meter is 100 centimeters), its pdf is given by the following table:

$S=1$	$S=2$	$S=3$	$S=4$
0.1	0.1	0.5	0.3

Then the expected height of someone wearing shoes in centimeters is

$$\mathbf{E}[100 \cdot H + S] = 100 \cdot \mathbf{E}[H] + \mathbf{E}[S] = 100 \cdot 1.755 + (0.1 \cdot 1 + 0.1 \cdot 2 + 0.5 \cdot 3 + 0.3 \cdot 4) = 175.5 + 3 = 178.5$$

Note how the linearity of expectation allowed us to decompose the expression $100 \cdot H + S$ into its components, and take the expectation of each one individually. This trick is immensely powerful when analyzing complex scenarios with many factors.

1.5 Variance

The *variance* of a random variable X describes how spread out it is. It is defined

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

The equivalence of those two common forms above uses that $\mathbf{E}[X]$ is a fixed scalar:

$$\mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] = \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

For any scalar $\alpha \in \mathbb{R}$, then $\mathbf{Var}[\alpha X] = \alpha^2 \mathbf{Var}[X]$.

Note that the variance does not have the same units as the random variable or the expectation, it is that unit squared. As such, we also often discuss the *standard deviation* $\sigma_X = \sqrt{\mathbf{Var}[X]}$.

Example: Variance

Consider again the random variable S for height added by a shoe:

$S=1$	$S=2$	$S=3$	$S=4$
0.1	0.1	0.5	0.3

Its expected value is $\mathbf{E}[S] = 3$ (a fixed scalar), and its variance is

$$\begin{aligned}\mathbf{Var}[S] &= 0.1 \cdot (1-3)^2 + 0.1 \cdot (2-3)^2 + 0.5 \cdot (3-3)^2 + 0.3 \cdot (4-3)^2 \\ &= 0.1 \cdot (-2)^2 + 0.1 \cdot (-1)^2 + 0 + 0.3(1)^2 = 0.4 + 0.1 + 0.3 = 0.8.\end{aligned}$$

Then the standard deviation is $\sigma_S = \sqrt{0.8} \approx 0.894$.

The *covariance* of two random variables X and Y is defined $\mathbf{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$. It measures how much these random variables vary in accordance with each other; that is, if both are consistently away from the mean at the same time (in the same direction), then the covariance is high.

1.6 Joint, Marginal, and Conditional Distributions

We now extend some of these concepts to more than one random variable. Consider two random variables X and Y. Their *joint pdf* is defined $f_{X,Y} : \Omega_X \times \Omega_Y \to [0, \infty]$ where for discrete random variables this is defined by the probability $f_{X,Y}(x, y) = \textbf{Pr}(X = x, Y = y)$. In this case, the domain of $f_{X,Y}$ is restricted so $f_{X,Y} \in [0, 1]$ and so $\sum_{x,y \in X \times Y} f_{X,Y}(x, y) = 1$.

Similarly, when $\Omega_X = \Omega_Y = \mathbb{R}$, the *joint cdf* is defined $F_{X,Y}(x, y) = \textbf{Pr}(X \leq x, Y \leq y)$. The *marginal cumulative distribution functions* of $F_{X,Y}$ are defined as $F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y)\mathrm{d}y$ and $F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x, y)\mathrm{d}x$.

Similarly, when Y is discrete, the *marginal pdf* is defined $f_X(x) = \sum_{y \in \Omega_Y} f_{X,Y}(x, y) = \sum_{y \in \Omega_Y} \textbf{Pr}(X = x, Y = y)$. When the random variables are continuous, we define $f_{X,Y}(x, y) = \frac{\mathrm{d}^2 F_{X,Y}(x,y)}{\mathrm{d}x\mathrm{d}y}$. And then the marginal pdf of X (when $\Omega_Y = \mathbb{R}$) is defined $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y$. Marginalizing removes the effect of a random variable (Y in the above definitions).

Now we can say random variables X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all x and y.

Then a *conditional distribution* of X given $Y = y$ is defined $f_{X|Y}(x \mid y) = f_{X,Y}(x, y)/f_Y(y)$ (given that $f_Y(y) \neq 0$).

Example: Marginal Distributions

Consider someone who randomly chooses his pants and shirt every day (a friend of mine actually did this in college – all clothes were in a pile, clean or dirty). Let P be a random variable for the color of pants, and S a random variable for the color of the shirt. Their joint probability is described by this table:

	S=green	S=red	S=blue
P=blue	0.3	0.1	0.2
P=white	0.05	0.2	0.15

Adding up along columns, the marginal distribution f_S for the color of the shirt is described by the following table:

S=green	S=red	S=blue
0.35	0.3	0.35

Isolating and renormalizing the middle "S=red" column, the conditional distribution $f_{P|S}(\cdot \mid S\text{= red})$ is described by the following table:

P=blue	P=white
$\frac{0.1}{0.3} = 0.3333$	$\frac{0.2}{0.3} = 0.6666$

Example: Gaussian Distribution

The Gaussian distribution is a d-variate distribution $N_d : \mathbb{R}^d \to \mathbb{R}$ that generalizes the one-dimensional normal distribution. The definition of the symmetric version (we will generalize to non-trivial covariance later on) depends on a mean $\mu \in \mathbb{R}^d$ and a variance σ^2. For any vector $\mathbb{R}^d$, it is defined

$$N_d(v) = \frac{1}{\sigma^d \sqrt{2\pi^d}} \exp\left(-\|v - \mu\|^2/(2\sigma^2)\right).$$

For the 2-dimensional case where $v = (v_x, v_y)$ and $\mu = (\mu_x, \mu_y)$, then this is defined

$$N_2(v) = \frac{1}{\sigma^2 \pi \sqrt{2}} \exp\left(-((v_x - \mu_x)^2 - (v_y - \mu_y)^2)/(2\sigma^2)\right).$$

A *magical* property about the Gaussian distribution is that all conditional versions of it are also Gaussian, of a lower dimension. For instance, in the two dimensional case $N_2(v_x \mid v_y = 1)$ is a 1-dimensional Gaussian, or a normal distribution. There are many other essential properties of the Gaussian that we will see throughout this text, including that it is invariant under all basis transformations and that it is the limiting distribution for central limit theorem bounds.