Chapter 1

A Review of the Principal Probability Distributions

In this chapter, we will bridge the gap between biological observations and mathematical theory. We will explore how 'Random Variables' represent biological traits and learn to use density and cumulative functions to predict outcomes. Finally, we will introduce the most important probability distributions that biologist needs to use.

1.1 1.1 A Random Variable

In the life sciences, even under strictly controlled and stable conditions, biological experiments rarely yield identical results. Probability theory provides the mathematical framework to study this inherent variability and the 'statistical regularity' found within natural phenomena.

1.1.1 1.1.1 The concept of random variables

Definition: Random Variable

Whenever we determine the height, weight, or age of an individual, the result is frequently referred to as a value of the respective variable. When the values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance, the variable is called a random variable.

Example: Random Variable

An example of a random variable is adult height. When a child is born, we cannot predict exactly his or her height at maturity. Attained adult height is the result of numerous genetic and environmental factors. Values resulting from measurement procedures are often referred to as observations or measurements.

Definition: Discrete Random Variable

Variables may be characterized further as to whether they are discrete or continuous. Since mathematically rigorous definitions of discrete and continuous variables are beyond the level of this chapter, we offer, instead, nonrigorous definitions and give an example of each. A discrete variable is characterized by gaps or interruptions in the values that it can assume. These gaps or interruptions indicate the absence of values between particular values that the variable can assume. Some examples illustrate the point. The number of daily admissions to a general hospital is a discrete random variable since the number of admissions each day must be represented by a whole number, such as 0, 1, 2, or 3. The number of admissions on a given day cannot be a number such as 1.5, 2.997, or 3.333. The number of decayed, missing, or filled teeth per child in an elementary school is another example of a discrete variable.

Definition: Continuous Random Variable

A continuous random variable does not possess the gaps or interruptions characteristic of a discrete random variable. A continuous random variable can assume any value within a specified relevant interval of values assumed by the variable. Examples of continuous variables include the various measurements that can be made on individuals such as height, weight, and skull circumference. No matter how close together the observed heights of two people, for example, we can, theoretically, find another person whose height falls somewhere in between. Because of the limitations of available measuring instruments, however, observations on variables that are inherently continuous are recorded as if they were discrete. Height, for example, is usually recorded to the nearest one-quarter, one-half, or whole inch, whereas, with a perfect measuring device, such a measurement could be made as precise as desired.

Example: Discrete vs. Continuous

Scenario: A marine biologist studies a population of sea turtles.

1. **Discrete R.V. (X):** The number of eggs in a nest. $X \in \{0, 1, 2, \ldots, n\}$. You cannot have 2.5 eggs.
2. **Continuous R.V. (Y):** The exact weight of a hatchling. Y could be $25.42g, 25.421g$, etc. It exists on a continuous scale.

1.1.2 1.1.2 The distribution of a random variable

1.1.2.1 Discrete random variable

Definition:

The probability distribution of a discrete random variable is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.

If X can take values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, the distribution satisfies:

- $p_i \geq 0$ for all i
- $\sum p_i = 1$

Example: Genotype Distribution

In a specific flower species, the probability of a seed being Red (x_1), Pink (x_2), or White (x_3) follows a Mendelian incomplete dominance pattern:

Color (x_i)	Red (AA)	Pink (Aa)	White (aa)
$P(X = x_i)$	0.25	0.50	0.25

Check: $0.25 + 0.50 + 0.25 = 1.0$. The conditions of a probability distribution are met.

1.1.2.2 Continuous random variable

Probability Density Function (PDF)

For a continuous r.v., we use a function $f(x)$ such that:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)dx$

Cumulative Distribution Function (CDF)

The CDF $F(x)$ expresses the probability that the variable X is **less than** a value x:

$$F(x) = P(X < x)$$

Example: Bacterial Growth

Let X be the time (hours) until a bacteria colony doubles. If the CDF is $F(x) = 1 - e^{-0.5x}$, find the probability it doubles within 2 hours.
Solution: $P(X \leq 2) = F(2) = 1 - e^{-0.5(2)} = 1 - e^{-1} \approx 0.632$.
There is a 63.2% chance the colony doubles within the first 2 hours.

1.1.3 Expectation (Mean) of a random variable

Mathematical Expectation μ

$$\begin{aligned} \textbf{Discrete: } \mu &= \sum x_i p_i \\ \textbf{Continuous: } \mu &= \int_{-\infty}^{+\infty} x f(x) dx \end{aligned}$$

Example: Expected Litter Size

If a rodent has a probability of 0.2 for 1 offspring, 0.5 for 2, and 0.3 for 3:
Solution: $E[X] = (1 \times 0.2) + (2 \times 0.5) + (3 \times 0.3) = 0.2 + 1.0 + 0.9 = 2.1$.
On average, we expect 2.1 offspring per litter.

1.1.4 Variance and Standard Deviation

Dispersion

$$\begin{aligned} \textbf{Variance: } \sigma^2(X) &= E[(X - E(X))^2] \\ \textbf{Standard Deviation: } \sigma(X) &= \sqrt{\sigma^2(X)} \end{aligned}$$

Example: Enzyme Activity

Two lab protocols produce the same mean enzyme activity ($\mu = 50$ units). Protocol A has $\sigma = 2$, Protocol B has $\sigma = 8$.
Conclusion: Protocol A is more reliable for biological experiments because it has a smaller standard deviation, meaning the results are more consistent.

1.1.5 Fractiles of a random variable

Definition: Quantiles/Fractiles

The quantile of order α (q_α) is the value such that $F(q_\alpha) = \alpha$.

Example: Biomass Fractile

In a study of tree diameters, the 0.90 fractile is found to be 45cm.
Meaning: 90% of the trees in this forest have a diameter less than or equal to 45cm.

1.2 1.2 Some common probability distributions

1.2.1 1.2.1 The Gaussian (Normal) Distribution

Definition:

We come now to the most important distribution in all of statistics—the normal distribution. The formula for this distribution was first published by Abraham De Moivre (1667–1754) on November 12, 1733. Many other mathematicians figure prominently in the history of the normal distribution, including Carl Friedrich Gauss (1777–1855). The distribution is frequently called the Gaussian distribution in recognition of his contributions. The normal density is given by

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

In the case where $\sigma = 1$ and $\mu = 0$, the distribution called standard normal distribution and his density function is: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The graph of the standard normal distribution produces the familiar bell-shaped curve shown in Figure bellow.

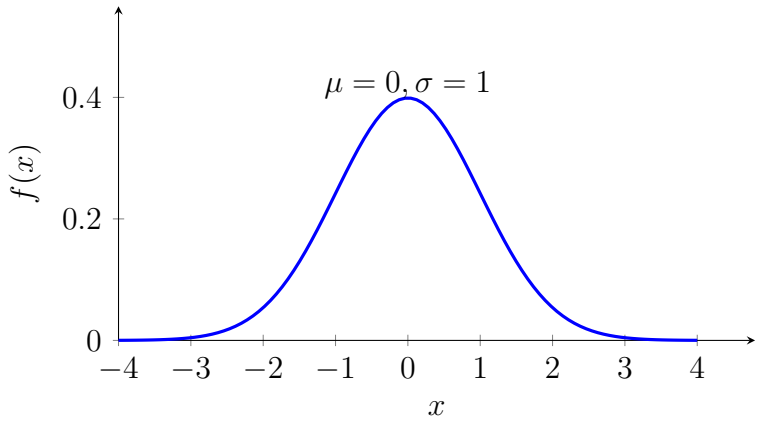


Figure 1.1: Standard Normal Distribution

Standardization (Z-score)

If $Y \sim N(\mu, \sigma^2)$, we transform it to $X \sim N(0, 1)$ using:

$$X = \frac{Y - \mu}{\sigma}$$

Example: Z-score Calculation

The height of a plant species follows $N(\mu = 20cm, \sigma = 4cm)$. What is the probability a plant is shorter than 26cm?
Step 1 (Standardize): $Z = \frac{26-20}{4} = \frac{6}{4} = 1.5$.
Step 2 (Table): Look up $P(Z < 1.5)$ in the Normal Table (Annexe).
Result: $\Phi(1.5) = 0.9332$. There is a 93.32% chance.

Characteristics of Normal Distribution

The following are some important characteristics of the normal distribution.

- It is symmetrical about its mean μ, As is shown in Figure, the curve on either side of is a mirror image of the other side.
- The mean, the median, and the mode are all equal.
- The total area under the curve above the x-axis is one square unit. This characteristic follows from the fact that the normal distribution is a probability distribution. Because of the symmetry already mentioned, 50 percent of the area is to the right of a perpendicular erected at the mean, and 50 percent is to the left.
- The normal distribution is completely determined by the parameters μ and σ. In other words, a different normal distribution is specified for each different value of μ and σ

Central Limit Theorem (CLT)

For a large sample ($n \geq 30$), the distribution of the sample mean $\bar{X}$ is approximately normal:

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

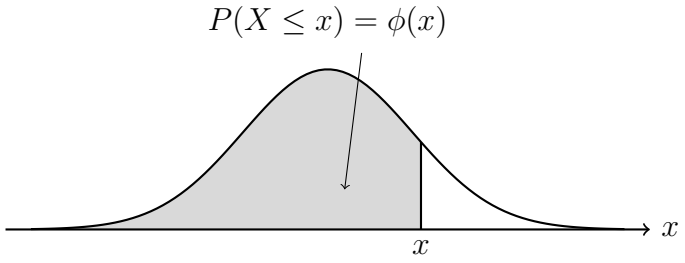
Normal Distribution Applications

Although its importance in the field of statistics is indisputable, one should realize that the normal distribution is not a law that is adhered to by all measurable characteristics occurring in nature. It is true, however, that many of these characteristics are approximately normally distributed. Consequently, even though no variable encountered in practice is precisely normally distributed, the normal distribution can be used to model the distribution of many variables that are of interest. Using the normal distribution as a model allows us to make useful probability statements about some variables much more conveniently than would be the case if some more complicated model had to be used. Human stature and human intelligence are frequently cited as examples of variables that are approximately normally distributed. On the other hand, many distributions relevant to the health field cannot be described adequately by a normal distribution. Whenever it is known that a random variable is approximately normally distributed, or when, in the

absence of complete knowledge, it is considered reasonable to make this assumption, the statistician is aided tremendously in his or her efforts to solve practical problems relative to this variable. Bear in mind, however, that "normal" in this context refers to the statistical properties of a set of data and in no way connotes normality in the sense of health or medical condition. There are several other reasons why the normal distribution is so important in statistics, and these will be considered in due time.

Distribution function of reduced centered normal distribution.

(Probability $\phi(x)$ of finding a value less than x)



x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.70	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.80	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.90	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.50	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.70	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.80	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.90	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.00	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.10	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.20	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.30	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.40	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.50	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.60	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.70	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.80	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.90	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.00	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.10	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.20	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.30	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.40	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.50	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.60	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.70	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.80	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.90	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997