Lecture 8: Basic Concepts in Probability

Probability theory makes extensive use of set notation and set operations. We begin this lecture with a brief review of some key concepts.

Set Notation and Set Operations [Bertsekas, 2008]

Recall that a *set* is a collection of objects which are *elements* of the set. If S is a set and x is an element of S we write $x \in S$ where the $\in$ is a symbol for "is an element of". If x is not an element of S we write $x \notin S$ where $\notin$ is a symbol for "is not an element of". A set can have no elements in which case it is called the emptyset, denoted by $\emptyset$. Sets can be specified in a variety of ways. If S contains a finite number of elements, say $x_1, x_2, \dots, x_n$ we write it as a list of the elements in braces:

$$S = \{x_1, x_2, \dots, x_n\} \tag{1}$$

Alternatively, we can consider a set of all x that have a certain property P and denote it by

$$S = \{x \mid x \text{ satisfies } P\} \tag{2}$$

where the symbol " $\mid$ " (or sometimes " $:$ ") is read as "such that". Occasionally the condition will include the phrase "for all" which is mathematically written using the symbol "$\forall$". If every element of a set S is also an element of the set T we say that S is a *subset* of T and we write $S \subset T$. If $S \subset T$ and $T \subset S$ the two sets are *equal* and we write $S = T$. It is also useful to consider a *universal set* Ω that contains all objects that could conceivably be of interest in a particular context (in probability theory Ω is the sample space of all possible outcomes of a random experiment). In the following we list some additional set notation and set-related operations:

- The *complement* of a set S with respect to the universal set Ω is $S^c = \{x \in \Omega \mid x \notin S\}$. The complement of a complement is the set itself, $(S^c)^c = S$, and the complement of the universal set is the emptyset, $\Omega^c = \emptyset$.

- The *powerset* of a set S is the set of all possible subsets (including the emptyset) and is denoted $\mathcal{P}(S)$. For example, suppose $S = \{a, b, c\}$, then

$$\mathcal{P}(S) = \{\emptyset, a, b, \{a, b\}, \{b, c\}, \{a, c\}, \{a, b, c\}\} \tag{3}$$

- The *union* of sets $S_1, \dots, S_n$ is denoted $S_1 \cup \cdots \cup S_n = \{x \in \Omega \mid x \in S_i \text{ for some } i = 1, \dots, n\}$. The union of the universal set with any subset is the universal set, $S \cup \Omega = \Omega$ for all $S \in \mathcal{P}(\Omega)$. The union of any set S with the emptyset $\emptyset$ is the set S, $S \cup \emptyset = S$. The union of a set with its complement is the universal set, $S \cup S^c = \Omega$ for all $S \in \mathcal{P}(\Omega)$.

- The *intersection* of sets $S_i, i = 1 \dots n$ is denoted by $S_1 \cap \cdots \cap S_n = \{x \in \Omega \mid x \in S_i \text{ for all } i = 1, \dots, n\}$. The intersection of the universal set with a set S is the set itself, $S \cap \Omega = S$. The intersection of a set with its complement is the emptyset, $S \cap S^c = \emptyset$.

- Sets S_1 and S_2 are *disjoint* if their intersection is the empty set, $S_1 \cap S_2 = \emptyset$. The difference $S_1 - S_2$ is the set of elements in S_1 but not in S_2, $S_1 - S_2 = \{x \in \Omega \mid x \in S_1, x \notin S_2\}$.

Many of these relations can be visualized using a Venn diagram as shown below.

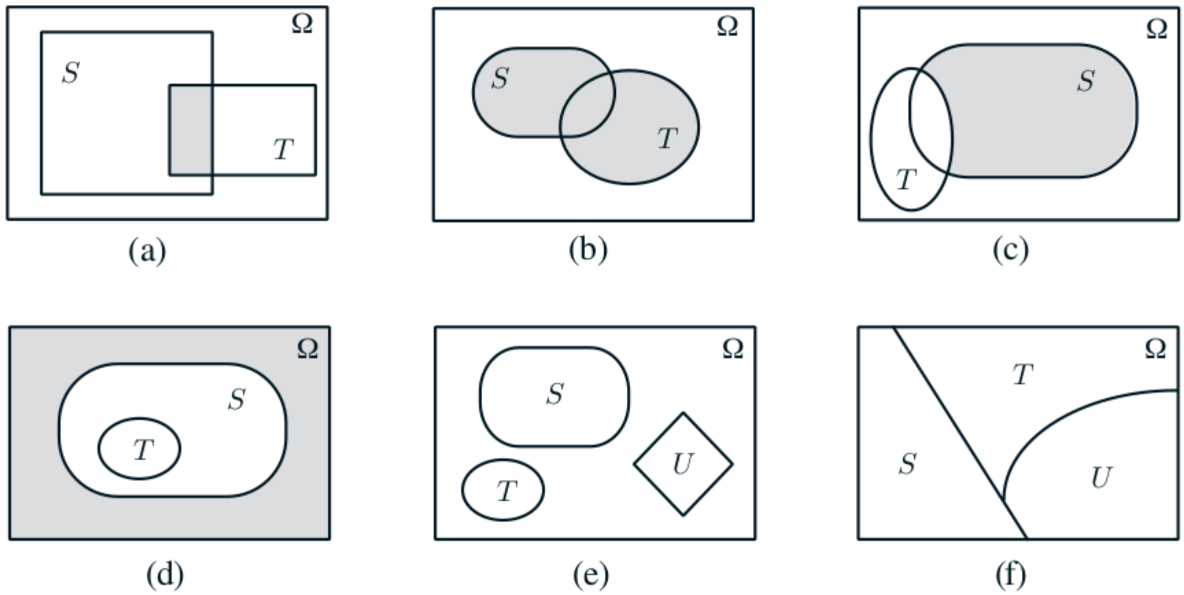


Figure 1.1: Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S. (e) The sets S, T, and U are disjoint. (f) The sets S, T, and U form a partition of the set Ω.

Figure 1: Image Source: [Bertsekas, 2008]

As a direct consequence of the above definitions and set operations we have the following properties concerning algebraic operations with sets (e.g., unions, intersections, complements). Let $S_1, S_2, S_3, \ldots, S_n$ and A all be sets that belong to the same universal set Ω.

- Associative Laws:

$$S_1 \cup (S_2 \cup S_3) = (S_1 \cup S_2) \cup S_3$$
$$S_1 \cap (S_2 \cap S_3) = (S_1 \cap S_2) \cap S_3$$

- Commutative Laws:

$$S_1 \cup S_2 = S_2 \cup S_1$$
$$S_1 \cap S_2 = S_2 \cap S_1$$

- Distributive Laws:

$$A \cap (\cup_{j=1}^{n} S_j) = \cup_{j=1}^{n} (A \cap S_j)$$
$$A \cup (\cap_{j=1}^{n} S_j) = \cap_{j=1}^{n} (A \cup S_j)$$

Probabilistic Models

A probabilsitic model is a mathematical description of an uncertain situation and has the following main components:

- A *random experiment*: a process that produces exactly one outcome, called a sample point ω, each time the process occurs. The set of all possible outcomes is the *sample space* Ω. Each element in the sample space is distinct (the elements are mutually exclusive and collectively exhaustive).
- A *probability law*: a rule, satisfying the axioms to be described next, that assigns a nonnegative number $P(A)$ to each subset of sample points. A collection of experiment outcomes (samples) $A \subseteq \mathcal{F}$ is called an *event* and $\mathcal{F}$ is the *event space*. For example, a valid event space is $\mathcal{F} = \mathcal{P}(\Omega)$.

A *simple event* contains only one sample point $A = \omega \in \Omega$, whereas a *composite event* contains multiple sample points, e.g., $B = \omega_1 \cup \omega_2 \subseteq \Omega$. (Note we also sometimes write the union of sample points as the event $B = \{\omega_1, \omega_2\}$.) These concepts are illustrated in the diagram below:

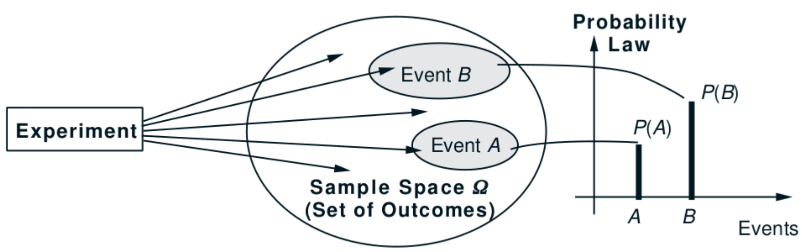


Figure 2: Image Source: [Bertsekas, 2008]

The probability law must satisfy the following *probability axioms*. Let Ω be a sample space with events $\{x_1, x_2, \cdots, x_n\}$. To every sample element $x \in \Omega$ (or event $x \subseteq \Omega$) the probability law must assign a number called the probability $P(x)$, that satisfies:

- (Nonegativity) $P(x) \geq 0$ for all $x \in \Omega$
- (Additivity) If x_i and x_j are disjoint events, then $P(\{x_i \cup x_j\}) = P(x_i) + P(x_j)$. Note this is the union of two elements so it gives the probability of either x_i or x_j. This is also denoted $P(x_i, x_j)$.
- (Normalization) The probability of the entire sample space is equal to 1.

$$P(\{x_1 \cup \cdots \cup x_n\}) = 1$$

Also, the set $A = \Omega$ is called a *certain event*. Conversely, the complement of event $A = \Omega$ is the emptyset, i.e., $\Omega^c = \emptyset$ and is called the *impossible event*. From these axioms we can deduce that, for some events in the event space $A, B \in \mathcal{F}$:

- If $A \subset B$, the $P(A) \leq P(B)$

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

When there is partial information available about the outcome of the random experiment we can use a *conditional probability law*. Suppose we know the outcome of the experiment is contained in event B then we wish to determine the likelihood it also belongs to event A. Introduce the conditional probability law as

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

where the " | " symbol in this context means "given". That is, the probability of event A given the outcome is in event B. One can confirm that this new probability law also satisfies the required probability axioms and thus they inherits all of the properties of a normal probability law. For example, recall that $P(A \cup C) \leq P(A) + P(C)$ and it is also true that

$$P(A \cup C|B) \leq P(A|B) + P(C|B) \ . \tag{4}$$

The conditional probability formula can be rearranged as $P(A \cap B) = P(A|B)P(B)$ and applying this iteratively gives a multiplication rule for several related events $A_1, A_2, \ldots, A_n$:

$$\begin{aligned}
P(A_1 \cap A_2 \cap \cdots \cap A_n) &= P(A_n|A_{n-1} \cap \cdots \cap A_1)P(A_{n-1} \cap \cdots \cap A_1) \\
&= P(A_n|A_{n-1} \cap \cdots \cap A_1)P(A_{n-1}|A_{n-2} \cap \cdots \cap A_1)P(A_{n-2} \cap \cdots \cap A_1) \\
&= P(A_n|A_{n-1} \cap \cdots \cap A_1)P(A_{n-1}|A_{n-2} \cap \cdots \cap A_1)P(A_{n-2}|A_{n-3} \cap \cdots \cap A_1) \\
&\quad \cdots \ P(A_2|A_1)P(A_1) \\
&= \prod_{i=2}^{n} P(A_i|A_{i-1} \cap \cdots \cap A_1)P(A_1)
\end{aligned}$$

Earlier in this lecture we alluded that the event space $\mathcal{F}$ can be the powerset of Ω. However, more preciesely, any $\mathcal{F}$ that is a *σ-algebra* suffices. A σ-algebra $\mathcal{F}$ is a collection of sets in Ω (not necessarily the powerset) that is "consistent" in the following sense:

1. If $A \in \mathcal{F}$ then its complement is also in $\mathcal{F}$, that is $A^c \in \mathcal{F}$
2. If $A_1, A_2 \in \mathcal{F}$ then their union is also in $\mathcal{F}$, that is $A_1 \cup A_2 \in \mathcal{F}$
3. The sample space itself is in $\mathcal{F}$, that is $\Omega \in \mathcal{F}$

Example: Consider a two-coin toss with possible outcomes of heads H or tails T for each coin. The sample space is $\Omega = \{HH, HT, TH, TT\}$. A valid σ-algebra (event space) is

$$\mathcal{F} = \{\emptyset, \Omega, \{TT\}, \{HT, TH, HH\}\} \ .$$

Indeed, other valid σ-algebras may be constructed with the sample space and note that the choice of $\mathcal{F}$ is not the same as the powerset of Ω.

For our purposes, we will not delve further into the measure-theoretic aspects of probability theory. The axiomatic definition of probability described earlier assumes a probability space defined by the triplet $(P, \Omega, \mathcal{F})$ which consists of a probability function, a sample space, and the event space sigma-algebra.

Random Variables

In many cases the probabilistic model introduced above has outcomes $\omega \in \Omega$ in the sample space that can be associated with numeric values (e.g., taking the temperature of an object with a noisy thermometer gives a sample space that consists of temperature values). A *random variable* (or r.v. for short) is a function that maps the outcomes of an experiment to the real number line $\mathbb{R} = (-\infty, \infty)$. For example, suppose we toss a coin 5 times, then the number of heads is an appropriate random variable. In this example, the random variable can only take on values of $\{1, 2, 3, 4, 5\}$. Random variables that only take on finite values are called *discrete random variables* whereas r.v.s that take on a continuum of values are called *continuous random variables*. Formally, a random variable can be considered a function $X(\omega) : \Omega \to T \subseteq \mathbb{R}$ that maps sample points ω to the real line $\mathbb{R} = (-\infty, \infty)$ (or a subset thereof). Random variables are typically denoted by capital letters, e.g., X. The range of X is the subset $T \subseteq \mathbb{R}$ of the real line that the random variable maps to: $T = X(\Omega) = \{X(\omega) : \omega \in \Omega\} \subseteq \mathbb{R}$. A r.v. is similar to a typical function, except that the argument (ω in our notation) lives in an abstract sample space. (For example, the sample space can be the outcome of a coin toss.) We are often interested in the probability that a r.v. takes on some particular value i.e., that $Z(\omega)$ is equal to some $z \in T$, where T is range of the r.v. This probability is denoted in shorthand as $P(Z = z)$ or $P(z)$. However, to be more precise, we could write

$$P(Z = z) = P(\omega \in \Omega : Z(\omega) = z) \tag{5}$$

to emphasize that in fact we are seeking the probability of the outcome in the sample space that maps through the random variable to $Z(\omega) = z$.

Discrete Random Variables

Consider a discrete r.v. $Z : \Omega \to T$, where $T = \{z_1, z_2, \dots, z_n\}$ is a finite set of values. A *probability mass function* (p.m.f) is a function $p_Z(z) = P(Z = z)$ that assigns to each value $z \in T$ a probability. (Note that p_Z assigns probabilities to random values whereas our earlier capital P assigns probabilities to sample points or events.) Summing the p.m.f. over the range of possible values adds to one

$$\sum_{z \in T} p_Z(z) = 1 \; .$$

The expectation of Z (also called the mean or average) is the sum of the values in T weighted by their probabilities

$$\begin{aligned} \mu = E[Z] &= z_1 p_Z(z_1) + \cdots + z_n p_Z(z_n) \\ &= \sum_{z \in T} z p_Z(z) \; . \end{aligned} \tag{6}$$

The second statistical moment of the r.v. is the expected value of Z^2,

$$\begin{aligned} E[Z^2] &= z_1^2 p_Z(z_1) + \cdots + z_n^2 p_Z(z_n) \\ &= \sum_{z \in T} z^2 p_Z(z) \; , \end{aligned} \tag{7}$$

and the Nth statistical moment is the expected value of Z^N

$$\begin{aligned} E[Z^N] &= z_1^N p_Z(z_1) + \cdots + z_n^N p_Z(z_n) \\ &= \sum_{z \in T} z^N p_Z(z) \ . \end{aligned} \tag{8}$$

When the moment is taken around the mean, i.e., when μ is subtracted from the r.v. in the expectation, we call this a *central moment*. The second central moment is the variance,

$$\begin{aligned} \sigma^2 = \mathrm{Var}(Z) &= E[(Z - E[Z])^2)] \\ &= E[(Z^2 - 2ZE[Z] + (E[Z])^2)] \\ &= E[Z^2] - 2ZE[Z] + (E[Z])^2] \\ &= E[Z^2] - 2E[Z]E[Z] + (E[Z])^2 \\ &= E[Z^2] - (E[Z])^2 \\ &= E[Z^2] - \mu^2 \end{aligned} \tag{9}$$

where σ is called the standard deviation. The *skew* of a random variable is a measured of the asymmetry of the pdf around its mean

$$\text{skew} = E[(X - \mu)^3]$$

In general, we also can define the following *moments* (sometimes called *point estimates*):

$$i\text{th moment of } X = E[(X)^i]$$
$$i\text{th central moment of } X = E[(X - \mu_x)^i]$$

It is also convenient to normalize the central moment by the corresponding power of the standard deviation. This corresponds to computing the statistical moments for a new *standardized r.v.*

$$\bar{Z} = \left(\frac{Z - \mu}{\sigma}\right) \ . \tag{10}$$

The standardized central moments are then:

$$E[\bar{Z}] = E\left[\left(\frac{Z - \mu}{\sigma}\right)\right] = \left[\left(\frac{E[Z] - \mu}{\sigma}\right)\right] = \left[\left(\frac{\mu - \mu}{\sigma}\right)\right] = 0 \quad \text{(standardized mean)} \tag{11}$$

$$E[\bar{Z}^2] = E\left[\left(\frac{Z - \mu}{\sigma}\right)^2\right] = \left(\frac{E[(Z - \mu)^2]}{\sigma^2}\right) = \left(\frac{\sigma^2}{\sigma^2}\right) = 1 \quad \text{(standardized variance)} \tag{12}$$

$$E[\bar{Z}^3] = E\left[\left(\frac{Z - \mu}{\sigma}\right)^3\right] \quad \text{(skew)} \tag{13}$$

$$E[\bar{Z}^4] = E\left[\left(\frac{Z - \mu}{\sigma}\right)^4\right] \quad \text{(kurtosis)} \ . \tag{14}$$

Continuous Random Variables

We can extend the concept of p.m.f. for discrete r.v.s to continuous r.v.s by introducing the *probability density function* (p.d.f.). For a continuous random variable, $Z : \Omega \to T$, where $T \subseteq \mathbb{R}$, the p.d.f. $f_Z(z)$ satisfies

$$P(Z \in B) = \int_B f_Z(s) ds \ , \tag{15}$$

where $B \subseteq T$. In other words, the probability that the continuous r.v. Z takes on a value in the interval B is equal to the integral (15). The function $f_Z(\cdot)$ is a curve over the real line that (to satisfy the probability axioms) must integrate to one

$$\int_{\mathbb{R}} f_Z(s)ds = 1 \; . \tag{16}$$

Note that a function may integrate to one even if it is greater than one at some points, i.e., the above requirement does not restrict $f_Z(\cdot) \leq 1$.

For the particular choice of the set $B_c(z) = \{s \in \mathbb{R} \; : \; s \leq z\}$, equivalently $B_c(z) = (-\infty, z]$ the integral (15) is

$$F_Z(z) = P(Z \in B(z)) = \int_{B_c(z)} f_Z(s)ds \tag{17}$$

$$= P(Z \leq z) = \int_{-\infty}^{z} f_Z(s)ds \tag{18}$$

and the function $F_Z(z)$ is called the *cumulative distribution function* (c.d.f.). Some properties obtained from this definition include:

$$\begin{aligned} F_Z(z) &\in [0,1] \\ F_Z(-\infty) &= 0 \\ F_Z(\infty) &= 1 \\ F_Z(a) &\leq F_Z(b) \quad \text{if} \quad a \leq b \\ P(a \leq Z \leq b) &= F_Z(b) - F_Z(a) \end{aligned}$$

The relationship of the c.d.f. with the p.d.f. is that

$$f_Z(s) = \left[\frac{dF_Z(\tau)}{d\tau} \right]_{\tau = s} \; .$$

Expected Value Operator for Continuous R.V.s and Functions

The expected value of any function $g(X)$ that depends on a random variable can be computed as follows

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)dx$$

where X is a continuous random variable and $p(x)$ is the probability density function of X (i.e., $p(x) = f_X(x)$). For $g(x) = X$ we obtain an expression for the mean/expected value:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx \tag{19}$$

For $g(X) = (X - \mu_x)^2$ we obtain the variance

$$\begin{aligned}
E[(X-\mu_x)^2] &= \int_{-\infty}^{\infty} (X-\mu_x)^2 p(x)dx \\
&= \int_{-\infty}^{\infty} X^2 p(x)dx - \int_{-\infty}^{\infty} 2X\mu_x p(x)dx + \int_{-\infty}^{\infty} \mu_x^2 p(x)dx \\
&= \int_{-\infty}^{\infty} X^2 p(x)dx - 2\mu_x \int_{-\infty}^{\infty} Xp(x)dx + \mu_x^2 \underbrace{\int_{-\infty}^{\infty} p(x)dx}_{=1} \\
&= E[X^2] - 2\mu_x E[X] + \mu_x^2 \\
&= E[X^2] - \mu_x^2
\end{aligned}$$

The definitions for central and standardized moments given above can also be obtained by replacing the summations with integrals, as needed.

Linearity of the Expected Value Operator

An operator L is said to be linear if

1. $L(f+g) = Lf + Lg$
2. $L(\alpha f) = \alpha Lf$

From the definition of the expected value (19)

$$\begin{aligned}
E[f(X)+g(X)] &= \int_{-\infty}^{\infty} (f(x)+g(x))p(x)dx && (20)\\
&= \int_{-\infty}^{\infty} f(x)p(x)dx + \int_{-\infty}^{\infty} g(x)p(x)dx && (21)\\
&= E[f(X)] + E[g(X)] && (22)
\end{aligned}$$

and

$$\begin{aligned}
E[\alpha g(X)] &= \int_{-\infty}^{\infty} \alpha g(x)p(x)dx && (23)\\
&= \alpha \int_{-\infty}^{\infty} g(x)p(x)dx && (24)\\
&= \alpha E[g(X)] && (25)
\end{aligned}$$

hence the expected value is linear. Values that are not random e.g., a constant k, remain unchaged by the expected value ($E[k] = k$).

Gaussian Random Variables

The p.d.f. of a Gaussian continuous r.v. has a *normal distribution* given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/2\sigma^2} , \tag{26}$$

where $\mu \in \mathbb{R}$ is the mean and $\sigma \in \mathbb{R}$ is the standard deviation. The mean μ determines where

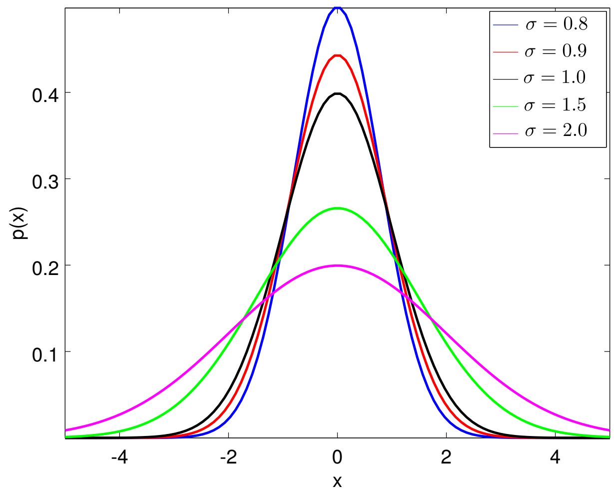


Figure 3: Gaussian probability distributions with a zero mean $\mu = 0$ and varying standard deviation σ

the Gaussian is centered and the variance σ^2 (or standard deviation σ) determines how spread out the distribution is i.e., the larger the standard deviation σ the wider the bell curve. A random variable Z that is normally distributed with mean μ and variance σ^2 is denoted by

$$Z \sim \mathcal{N}(\mu, \sigma^2) \ .$$

For example, the blue curve in Fig. 3 represents the p.d.f (26) of $Z \sim \mathcal{N}(0, 0.8^2)$.

As evident from (26), the Gaussian p.d.f mean and variance fully define the shape of the p.d.f curve. For a Gaussian r.v. Z, the standardized moments are

$$E[\bar{Z}] = 0 \quad \text{(standardized mean for Gaussian r.v.)} \tag{27}$$

$$E[\bar{Z}^2] = 1 \quad \text{(standardized variance for Gaussian r.v.)} \tag{28}$$

$$E[\bar{Z}^3] = 0 \quad \text{(skew for Gaussian r.v.)} \tag{29}$$

$$E[\bar{Z}^4] = 3 \quad \text{(kurtosis for Gaussian r.v.)} \tag{30}$$

As discussed previously, the probability that a random variable Z takes on a value in some set (e.g., $B_I(a,b) = \{s \in \mathbb{R} \ : \ a \leq s \leq b\}$) is found by integrating the p.d.f over the set:

$$P[B_I(a,b)] = \int_a^b f_Z(s) ds \ . \tag{31}$$

If Z is Gaussian, then (31) requires integrating (26) which has no closed-form expression. However, some values of the integral (31) can be found in a *cumulative probability table* that was computed numerically. For example, the probability of Z taking on a value within one standard deviation of the mean is about 68%:

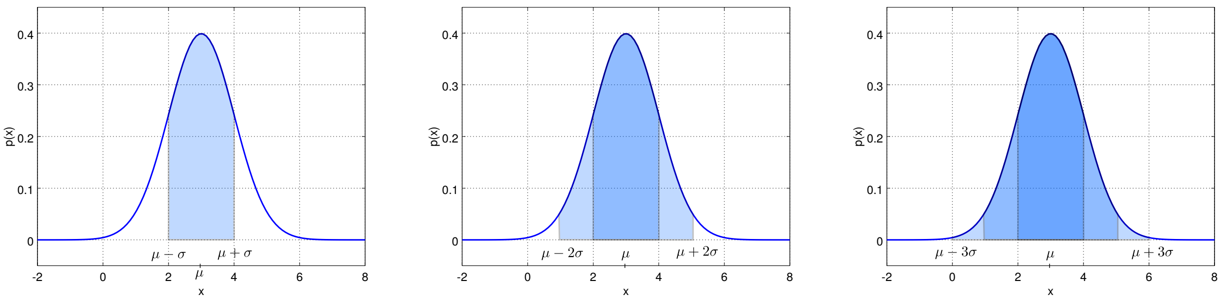
$$P[B_I(\mu - \sigma, \mu + \sigma)] = 0.68 \ . \tag{32}$$

Similarly, the probability of X taking on a value within two or three standard deviations of the mean is:

$$P[B_I(\mu - 2\sigma, \mu + 2\sigma)] = 0.95 \tag{33}$$

$$P[B_I(\mu - 3\sigma, \mu + 3\sigma)] = 0.997 \tag{34}$$

Refer to Fig. 4 for a graphical representation.



(a) 68% of the time the random number is within $\mu \pm \sigma$

(b) 95% of the time the random number within $\mu \pm 2\sigma$

(c) 99.7% of the time the random number is within $\mu \pm 3\sigma$

Figure 4: Probability distribution for the random number $X \sim \mathcal{N}(3, 1^2)$. The mean is $\mu = 3$ and the standard deviation is $\sigma = 1$.

References

[Bertsekas, 2008] Bertsekas, D. (2008). *Introduction to Probability*. Athena Scientific.