

# A Statistics Primer: Rules of Probability and Probability Distributions

Vincent A. Voelz

E-mail: [vvoelz@gmail.com](mailto:vvoelz@gmail.com)

*Math Bio Boot Camp 2006*

(adapted from BP203 2002 lecture notes)

University of California at San Francisco, San Francisco, CA 94143

August 29, 2006

## Contents

<b>1</b>	<b>Probability</b>	<b>2</b>
1.1	Rules of Probability . . . . .	3
1.1.1	Set notation. . . . .	3
1.1.2	Probability identities. . . . .	4
1.2	Bayes' Theorem . . . . .	6
1.3	Permutations and Combinations . . . . .	7
1.4	Expectation value and variance . . . . .	7
<b>2</b>	<b>Discrete Probability Distributions</b>	<b>8</b>
2.1	The Binomial Distribution . . . . .	8
2.2	Properties of the Binomial Distribution . . . . .	9
2.2.1	Bernoulli variables . . . . .	9
2.3	The Poisson Distribution . . . . .	10

1	PROBABILITY	2
2.4	The Multinomial Distribution . . . . .	11
2.5	The Hypergeometric Distribution . . . . .	12
<b>3</b>	<b>Continuous Probability Distributions</b>	<b>14</b>
3.1	Manipulating continuous distributions . . . . .	15
3.1.1	The expectation value . . . . .	15
3.1.2	The moment-generating function . . . . .	16
3.2	The Normal Distribution . . . . .	16
3.2.1	Normalization . . . . .	17
3.2.2	Properties of the normal distribution . . . . .	18
3.3	Changes of Variable . . . . .	19
3.3.1	Finding a coordinate transformation . . . . .	19
3.3.2	Example: generating normal variables . . . . .	19
3.3.3	Distribution transformations under changes of variables . . . . .	20
3.4	The Central Limit Theorem . . . . .	22
3.5	The $\chi^2$ Distribution . . . . .	22
3.6	The $t$ -distribution . . . . .	24

# 1 Probability

Statistical methods attempt to quantify the *likelihood* of certain events occurring. There are two philosophical interpretations of what we mean by "probability".

**The frequency interpretation.** When we toss a coin, we say that there is  $p(H) = 1/2$  probability that it lands heads, and  $p(T) = 1/2$  probability that it lands tails. The frequency interpretation of this is that if we tossed the coin  $N$  times, then on average we would find that as  $N$  increases,  $N/2$  would be heads and  $N/2$  would be tails. This interpretation restrains our predictions by the available data. We can't say anything

about the outcome of a *single* toss. By this interpretation, the probability of an event  $x$  is:

$$p(x) = \frac{N_x}{N} \quad (1)$$

where  $N$  is the total number of trials, and  $N_x$  is the number of trials where event  $x$  is observed.

**The sample space interpretation.** When the meteorologist says there will be a 40% chance of rain today, according to the frequency interpretation this has no meaning; obviously, there is only one "today", and when it is over we shall not have another. The sample space interpretation is that there is a *sample space* of all possible events, each of which are equally likely. For instance, let's say we toss a coin three times in a row, we have eight equally likely possibilities in our sample space:

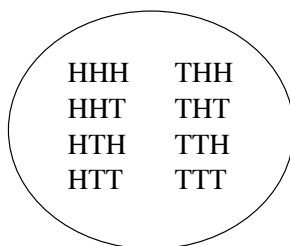


Figure 1: The sample space of three coin tosses. H is heads, and T tails.

By the sample space interpretation, the probability of an event  $x$  is:

$$p(x) = \frac{\text{the number of ways } x \text{ can occur}}{\text{the number of possible outcomes}} \quad (2)$$

For example, the probability of getting the sequence HTH is  $p(\text{HTH}) = 1/8$ . The probability of getting exactly two heads,  $p(\text{getting exactly 2 heads}) = 3/8$ , because there are three outcomes (HHT, HTH, THH) out of eight that have exactly two heads.

## 1.1 Rules of Probability

The rules of probability come directly from thinking about events as *sets* of outcome elements in the sample space.

### 1.1.1 Set notation.

Let's denote the sample space as the set  $\Omega$ . Recall some basic set notation:

$$x \cup y = \text{the union of } x \text{ and } y \quad (3)$$

$$x \cap y = \text{the intersection of } x \text{ and } y \quad (4)$$

Also, recall some basic properties of sets: If we have two sets  $x$  and  $y$  that share elements, then

$$N(x \cup y) = N(x) + N(y) - N(x \cap y) \quad (5)$$

where  $N(x)$  denotes the number of elements in set  $x$ . This can be easily confirmed by counting; the elements  $(x \cap y)$  shared between  $x$  and  $y$  are counted twice and must be subtracted off.

If  $x$  and  $y$  are *disjoint*, meaning  $N(x \cap y) = 0$ , then it follows that:

$$N(x \cup y) = N(x) + N(y) \quad (6)$$

In general, the probability function  $p()$  can be thought of providing a mapping from the sample space to the interval  $[0,1]$  by counting elements, and normalizing by the total number of elements in the sample space,  $N(\Omega)$ .

$$p(x) = \frac{N(x)}{N(\Omega)} \quad (7)$$

### 1.1.2 Probability identities.

**Inclusive and exclusive probabilities.** If  $x$  and  $y$  are any two events in the outcome space, then the probability of either  $x$  OR  $y$  occurring is

$$p(x \cup y) = p(x) + p(y) - p(x \cap y) \quad (8)$$

If  $x$  and  $y$  are *mutually exclusive* events, such that if event  $x$  occurs,  $y$  will not, and vice versa, then

$$p(x \cup y) = p(x) + p(y) \quad (9)$$

**Probability Conservation.** If  $\{x_i\}$  is a collection of all possible disjoint (i.e. mutually exclusive) outcomes, then

$$\sum_i p(x_i) = 1 \quad (10)$$

**Conditional probabilities.** The *conditional* probability of observing event  $x$  given that event  $y$  has already been observed, is denoted  $p(x | y)$ . By the frequency interpretation,

$$p(x | y) = \frac{N(x \cap y)}{N(y)} = \frac{N(x \cap y)/N(\Omega)}{N(y)/N(\Omega)} = \frac{p(x \cap y)}{p(y)} \quad (11)$$

**Joint probabilities.** If  $x$  and  $y$  are any two events in the outcome space, then the *joint* probability of  $x$  AND  $y$  occurring simultaneously is

$$p(x \cap y) \quad (12)$$

**Independence.** Two outcomes are said to be *independent* if

$$p(x \cap y) = p(x)p(y) \quad (13)$$

This is equivalent to saying that the conditional probability  $p(x | y) = p(x)$ :

$$p(x | y) = \frac{p(x \cap y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x) \quad (14)$$

**Example 1.** Let us now return to our example of tossing a coin three times in a row. Let the event  $x$  be "the first coin is H" and event  $y$  be "the second coin is H". What is the probability of  $p(x \text{ and } y)$ ?

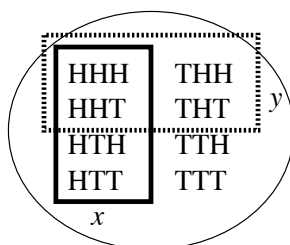


Figure 2: The intersection of two sample spaces:  $x$  is the set where the first coin is H, and  $y$  is the set where the second coin is H.

**Answer:** The first coin toss should be completely independent from the second coin toss. Therefore,  $p(x \text{ and } y) = p(x)p(y) = (1/2)(1/2) = 1/4$ .

**Example 2.** Let the event  $x$  be "the first coin is H" and event  $y$  be "the first coin is T". What is the probability of  $p(x \text{ or } y)$ ? (Figure 3).

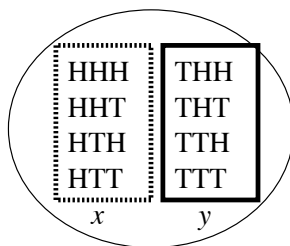


Figure 3:

**Answer:** The first coin can either be H or T, and these two events are mutually exclusive. Therefore,  $p(x \text{ or } y) = p(x) + p(y) = (1/2) + (1/2) = 1$ .

## 1.2 Bayes' Theorem

Bayes' Theorem is the basis for Bayesian statistics, a powerful way of using known information about prior observed events to construct probabilistic models. We will be saying more about these methods later in the course, and will here simply state the theorem, which can be derived from the rules of probability shown above.

Let  $A$  and  $B_1, B_2, \dots, B_n$  be events where  $B_i$  are *disjoint* and cover the whole sample space, (i.e.  $\bigcup_{i=1}^n B_i = \Omega$ ). Let  $p(B_i) > 0$  for all  $i$ . Then

$$p(B_j | A) = \frac{p(A | B_j)p(B_j)}{\sum_{i=1}^n p(A | B_i)p(B_i)} \quad (15)$$

We leave the derivation as an exercise for the reader.

**Example: Transmitting a message through a noisy channel.** Consider a very simple communication system where only three symbols,  $w_1, w_2$ , and  $w_3$  are sent to a receiver to be interpreted as messages  $y_1, y_2$ , and  $y_3$ , respectively. (For instance, say the symbols are pulses of voltage sent down a telegraph wire to some kindly Western Union operator, and the symbols are letters.) Suppose there is a noisy connection, and there are probabilities  $p(y_j | w_i)$  that symbol  $w_i$  is misinterpreted as symbol  $y_j$ .

Suppose that each symbol  $w_i$  is sent with equal probability,  $p(w_i) = 1/3$ . **Question:** If the receiver interprets some symbol as message  $y_j$ , what's the probability that the *actual* symbol sent was  $w_i$ ?

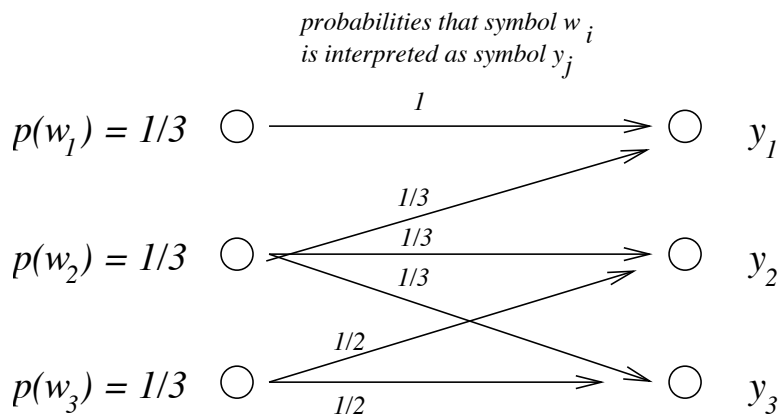


Figure 4: The cross-over probabilities that symbol  $y$  was received while symbol  $w$  was actually sent.

**Answer:** According to Bayes Theorem, the probability that message  $y_j$  actually came from symbol  $w_i$  is

$$p(w_i | y_j) = \frac{p(y_j | w_i)p(w_i)}{\sum_{i=1}^n p(y_j | w_i)p(w_i)} \quad (16)$$

So if the receiver interprets a symbol as  $y_2$ , there's a chance that in *actuality* symbol  $w_3$  was sent:

$$p(w_3 | y_2) = \frac{(1/2)(1/3)}{(1/2)(1) + (1/3)(1/3) + (0)(1/3)} = 3/5 \quad (17)$$

### 1.3 Permutations and Combinations

**Permutations.** Suppose you have  $N$  gumballs, each uniquely numbered 1 to  $N$ . You put them all in a hat, and draw all  $N$  of them out one by one. How many possible gum ball orderings are there? Well, on the first draw, there are  $N$  gumballs to choose from. On the second draw, there are  $N-1$  remaining possibilities, and so on. So the total number of possible orderings are

$$N! = N(N-1)(N-2)\dots(2)(1) \quad (18)$$

where  $N!$  is the *factorial* function. When spoken,  $N!$  is read "N factorial". There are  $N!$  different *permutations*, or orderings, of  $N$  distinct objects.

**Combinations.** Now suppose that our objects (gumballs) are non-distinct (not numbered) except that some are white and some are black. How many distinguishable combinations are there for a collection of  $m$  black gumballs and  $N - m$  white gumballs?

Like before, there are  $N!$  permutations of the entire ensemble of gumballs, but without the gumballs being distinguishable (no numbers) there will be many orderings will look the same to us when we draw them out. For example, if all the gumballs are black, then there is only a single distinguishable combination. Since the combination of the gumballs looks the same for all rearrangements of the  $m$  black balls and the  $N - m$  white balls,  $N!$  overestimates the number of distinguishable combinations by factors of  $m!$  and  $(N - m)!$  respectively:

$$\text{number of combinations} = \frac{N!}{m!(N-m)!} = \binom{N}{m} \quad (19)$$

where  $\binom{N}{m}$  is the *binomial* formula, read "n choose m".  $\binom{N}{m}$  is the number of combinations of distinct subgroups of  $m$  taken from  $N$  objects.

### 1.4 Expectation value and variance

Here we introduce expectation value and variance, two quantities that are often used in statistics to characterize distributions.

**Expectation value** The *expectation value* of a distribution is denoted  $E(x)$ . It can be thought of as the "weighted average" of all possible  $x$  across the probability distribution  $P(m)$ .

$$E(x) = \sum_{\text{all } x} xP(x) \quad (20)$$

In this sense,  $E(x)$  is the *mean* outcome of the distribution of outcomes  $P(x)$ . This notation is very handy as convenient shorthand, and we shall see it throughout the course.

Note that the expectation value is not necessarily the most *probable* outcome, which is

$$\max_{\text{all } x} P(x) \quad (21)$$

**Variance** The *variance* of a distribution is defined

$$\text{var}(x) = \sum_{\text{all } x} (x - E(x))^2 P(x) \quad (22)$$

It is a measure of the "width" of the distribution, i.e. how much the values of the distribution stray from the mean. The variance is often denoted as  $\sigma^2$ , and  $\sigma$  in turn is called the *standard deviation*.

It can be shown that above expression simplifies to:

$$\text{var}(x) = E(x^2) - E(x)E(x) \quad (23)$$

The derivation is an exercise left to the reader.

## 2 Discrete Probability Distributions

### 2.1 The Binomial Distribution

Suppose we toss a coin  $N$  times, and record how many total heads and tails we get. What is the probability  $P(m)$  that we get  $m$  heads?

For any one particular sequence (i.e. ordering) of H's and T's, the probability is  $(1/2)^N$ , because each toss is independent from the others. Furthermore, there are  $\binom{N}{m}$  such sequences with the right numbers of H's and T's. So the probability is

$$P(m) = \binom{N}{m} (1/2)^N \quad (24)$$

Now suppose we have a *biased* coin, such that the probability of getting heads is  $p$  and the probability of getting tails is  $(1 - p)$ . Now the probability of getting a sequence of  $m$  H's and  $(1 - m)$  T's is  $p^m(1 - p)^{N-m}$ , so the probability is

$$P(m) = \binom{N}{m} p^m (1 - p)^{N-m} \quad (25)$$

This is the *binomial distribution*. It is plotted below for  $N = 20$  and  $p = 0.5$ ,  $p = 0.2$ .

Figure 5: The binomial distribution for  $N = 20$ ,  $p = 0.5$

Figure 6: The binomial distribution for  $N = 20$ ,  $p = 0.2$

The binomial distribution is a function that gives up a probability  $P(m)$  for any  $m$ . It is a *distribution* in the sense that while the total probability of all events must equal one, the probability is distributed across many events with various values of  $m$ .

## 2.2 Properties of the Binomial Distribution

What is the expectation value and variance of the binomial distribution? From our definition of expectation value, we can write down the expression:

$$E(x) = \sum_{\text{all } x} x \binom{N}{x} p^x (1-p)^{N-x} \quad (26)$$

Since this is a rather clumsy expression, we will use another technique to find the mean and the variance.

### 2.2.1 Bernoulli variables

Consider a random Bernoulli variable  $x_i$ , defined probabilistically:

$$x_i = \begin{cases} 1 & \text{with (H) probability } p, \\ 0 & \text{with (T) probability } 1-p \end{cases} \quad (27)$$

We'll define the variable  $x$  to be:

$$x = \sum_{i=1}^N x_i \quad (28)$$

Thus  $x$  reports the number of heads obtained after tossing the coin  $N$  times. So if we want to know the expectation value  $E(x)$ , we can calculate the mean value of  $x$  (denoted  $\langle x \rangle$ ) like so:

$$E(x) = \langle x \rangle = \left\langle \sum_{i=1}^N x_i \right\rangle = \sum_{i=1}^N \langle x_i \rangle = N \langle x_i \rangle \quad (29)$$

and  $\langle x \rangle$  is just the probability-weighted average  $(1)(p) + (0)(1-p) = p$ , so:

$$E(x) = Np \quad (30)$$

Similarly, we can calculate the variance:

$$\begin{aligned}
 \text{var}(x) &= \sum_{i=1}^N \text{var}(x_i) = N \text{var}(x_i) \\
 &= N(\langle x_i^2 \rangle - (\langle x_i \rangle)^2) \\
 &= N(p - p^2) \\
 \text{var}(x) &= Np(1 - p)
 \end{aligned} \tag{31}$$

### 2.3 The Poisson Distribution

The Poisson distribution is the limiting case of the binomial distribution as  $p \rightarrow 0, N \rightarrow \infty$ , such that their product is a finite number  $Np = \lambda$ .

To illustrate this, consider the example of radioactive decay. A radioactive particle has a half life of  $\tau$ . The probability  $p$  of decaying in a small time interval  $dt$  is thus

$$p = \frac{dt}{\tau} \tag{32}$$

where  $\tau \gg dt$ . Let's say we start with  $N$  radioactive particles. Then after time  $dt$  has passed, this number has changed:

$$\begin{aligned}
 dN &= -N \frac{dt}{\tau} \\
 \frac{dN}{dt} &= \frac{-N}{\tau}
 \end{aligned} \tag{33}$$

From this equation we get our familiar expression for exponential rate of decay:  $N(t) = N_0 e^{-t/\tau}$ , where  $N_0$  is the number of particles at time zero.

If we set up a Geiger counter to count individual decays, we could record the number of counts in a interval  $\tau$ , and compile a histogram of the results. Alternatively, we could use the binomial distribution to consider the distribution of a large number ( $N$ ) of "passing moments" ( $dt$ ), each with an extremely rare chance of decay ( $p$ ), where  $Np = \lambda$ .

The number of decays observed,  $m$ , will be very small compared to the number of "passing moments",  $dt$ . Thus, we can make the approximations

$$\begin{aligned}
 N - m &\approx N \\
 \frac{N!}{(N - m)!} &\approx N^m
 \end{aligned} \tag{34}$$

Then the binomial distribution becomes

$$P(m) \approx \frac{N^m p^m}{m!} (1-p)^N \quad (35)$$

and because  $\lim_{p \rightarrow 0} (1-p)^{1/p} = e^{-1}$ ,

$$P(m) = \frac{\lambda^m}{m!} e^{-\lambda} \quad (36)$$

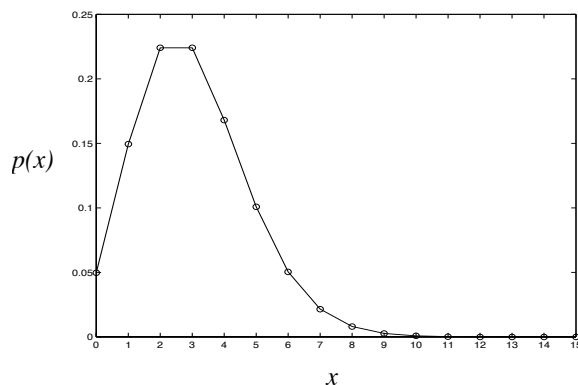


Figure 7: The Poisson distribution, for  $\lambda = 3$

**Mean and variance.** What is the mean and variance of the Poisson distribution? Since it is merely a limiting case of the binomial distribution,

$$E(x) = Np = \lambda \quad (37)$$

$$\text{var}(x) = Np(1-p) \rightarrow Np = \lambda \quad (38)$$

Apparently, the mean and the variance are the same value, so in a Geiger counter experiment, say, we would expect the number of counts to fluctuate with standard deviation  $\sigma = \sqrt{\lambda}$  about the mean value  $\lambda$ .

## 2.4 The Multinomial Distribution

The multinomial distribution is a natural extension of the binomial distribution. Here, instead of having  $N$  trials, each with only 2 outcomes possible (H or T), the multinomial distribution considers  $S$  possible outcomes, each with probabilities  $p_1, p_2, \dots, p_S$  of occurring.

An example would be a biased 6-sided die, with probabilities  $p_1, p_2, \dots, p_6$ . The probability of observing a set of counts for each outcome  $n_1, n_2, \dots, n_S$ , (where, of course,  $\sum_i n_i = 1$ ) is

$$P(n_1, n_2, \dots, n_S) = \frac{N!}{n_1! n_2! \dots n_S!} p_1^{n_1} p_2^{n_2} \dots p_S^{n_S} \quad (39)$$

The derivation uses arguments identical to the derivation of the binomial distribution, and we will leave this as an exercise for the reader.

## 2.5 The Hypergeometric Distribution

Suppose we wish to choose from a population which contains only two types of objects. This is akin to the example of picking black and white gumballs from a jar. Imagine there are  $m_1$  type 1 objects and  $m_2$  type 2 objects and we reach in our jar and pull out  $N$  objects. What is the probability of selecting  $x$  objects from population  $m_1$  and  $N - x$  objects from population  $m_2$  (thus making  $N$  total selections)?

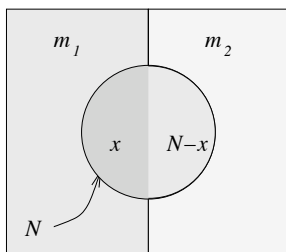


Figure 8: A graphical representation of the hypergeometric distribution.

Well, the  $x$  members can be chosen in  $\binom{m_1}{x}$  ways and the  $N - x$  members can be chosen in  $\binom{m_2}{N-x}$ . In total, there are  $\binom{m_1+m_2}{N}$  selections possible and since  $p(x) = \frac{N_x}{N}$

$$p(x) = \frac{\binom{m_1}{x} \binom{m_2}{N-x}}{\binom{m_1+m_2}{N}}$$

We can expand this expression to write:

$$P(x, m_1, m_2) = \frac{m_1! m_2! N! (m_1 + m_2 - N)}{(m_1 - x)! x! (m_2 - (N - x))! (N - x)! (m_1 + m_2)!} \quad (40)$$

Note that when  $x = N/2$  this equation is symmetric with respect to  $m_1$  and  $m_2$ .

**Example: The card game Hearts** In the card game of Hearts, 4 players are each dealt a hand of 13 cards. In a hand, an aggressive player may try to "shoot the moon", which requires having as many hearts as possible. Suppose we wish to calculate the probability of being dealt 7 hearts. Since we have two populations, hearts and non-hearts, we may apply the hypergeometric distribution equation, with  $m_1 = 13$ ,  $m_2 = 39$ ,  $x = 7$ , and  $N - x = 6$ .

$$p(7 \text{ hearts}) = \frac{\binom{13}{7} \binom{39}{6}}{\binom{52}{13}} \approx .0082$$

**Example: Correlations of two regulatory motifs in a set of promoter sequences** Let's say we want to examine the frequencies of certain sequence motifs within a set of regulatory promoters. Each of  $N$  promoter genes either contains motif 1, motif 2, or both motifs. Let's say we discover that  $m_1$  promoters have motif 1,  $m_2$  promoters have motif 2, and  $x$  promoters have both motifs. Are these overlaps simply a result of random chance?

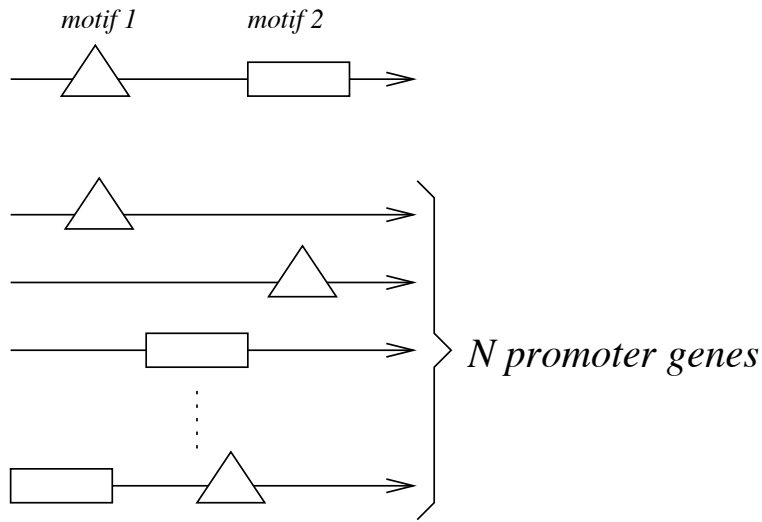


Figure 9: Regulatory motifs in a set of promoter sequences.

Using the hypergeometric distribution, we can calculate the probability distribution of  $(m_1, m_2, x)$  and assess how probable the measured result is simply by random chance. If it's not very probable, then perhaps there is some functional reason for the high occurrence of overlap.

### 3 Continuous Probability Distributions

Previously, we have been considering the probability of discrete events. For example, the probability of rolling a die and getting  $x$  is graphed below.

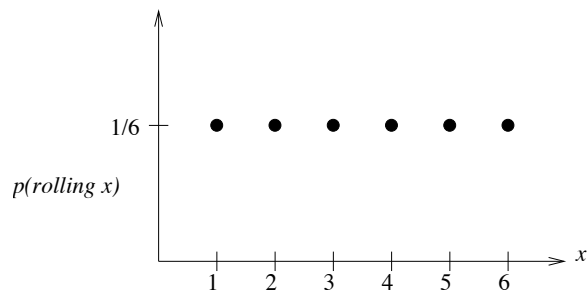


Figure 10:

From the graph, we can see that the probability of rolling a "3" is  $p(3) = 1/6$ .

Now consider the case where a random variable is drawn from a pool of *continuous* numbers, say, any real number  $x \in [0, 6]$ . A graph of this continuous probability is shown below.

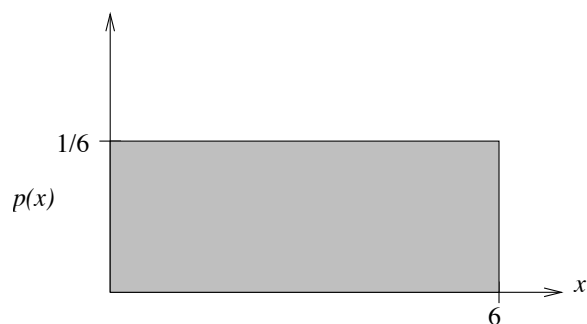


Figure 11: Continuous distribution of real numbers drawn from  $[0,6]$

In this case, the quantity  $p(3.0)$  has no meaning, because there are *infinitely* many real numbers in the interval  $[0, 6]$ , so the probability of drawing 3.0 exactly is infinitely improbable!

For continuous distributions, then, we must think in terms of differentials. We say that  $p(x)dx$  is the probability that the number drawn is between  $x$  and  $x + dx$ .

Like discrete probabilities, the total summed probability over all outcomes must be *normalized* to 1. Whereas for discrete distributions, we require that  $\sum_i p_i = 1$ ; for continuous distributions, we require that

$$\int_{-\infty}^{\infty} p(x)dx = 1 \quad (41)$$

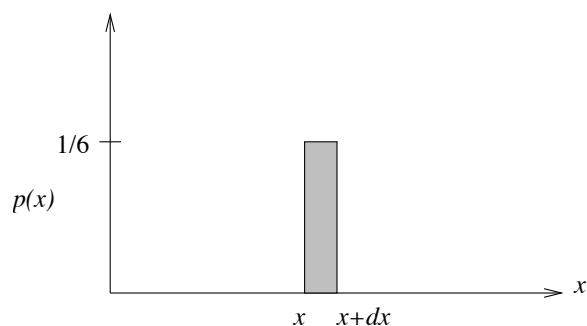


Figure 12:

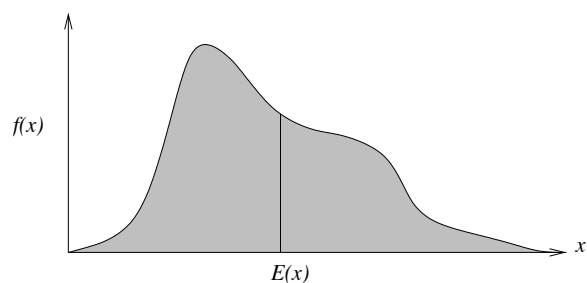
In the above graph of  $p(x)$ , we can easily see that the area under the curve,  $\int p(x)dx = 1$ .

### 3.1 Manipulating continuous distributions

Here we would like to introduce some tools and further nomenclature regarding continuous distributions.

#### 3.1.1 The expectation value

Consider some arbitrary continuous distribution  $f(x)$ :

Figure 13: The continuous distribution  $f(x)$  and its expectation value  $E(x)$ 

Like in the discrete case, the *expectation value* of a continuous distribution is denoted  $E(x)$ . It can be thought of as the "weighted average" of all possible  $x$  across the probability distribution.

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx \quad (42)$$

In this sense,  $E(x)$  is the *mean* of the distribution  $f(x)$ .

### 3.1.2 The moment-generating function

In general, the  $m$ th *moment* of a distribution is defined as

$$E(x^m) = \int_{-\infty}^{\infty} x^m f(x) dx \quad (43)$$

The expectation value  $E(x)$ , the *mean*, is the first moment of  $f(x)$ . The *variance* can be expressed in terms of the first and second moments:

$$\text{variance} = E(x^2) - E(x)E(x) \quad (44)$$

A useful piece of mathematical machinery is the *moment-generating function*, defined as

$$\psi^m(t) = \left(\frac{d}{dt}\right)^m \int_{-\infty}^{\infty} e^{tx} f(x) dx \quad (45)$$

To find the  $m$ th moment of  $f(x)$ , we evaluate the moment-generating function at  $t=0$

$$\psi^m(0) = E(x^m) \quad (46)$$

As an interesting side note, it turns out that if you know all of the moments of a distribution, you can specify the distribution uniquely. (This is somewhat reminiscent of, say, a Taylor's series expansion, where any function can be described by its full expansion of polynomial coefficients.)

## 3.2 The Normal Distribution

The *normal* distribution is defined as:

$$N(\mu, \sigma) = A e^{-(x-\mu)^2/2\sigma^2} \quad (47)$$

where  $\mu$  is the mean of the distribution,  $\sigma^2$  is the variance, and  $A$  is a *normalization constant*, chosen such that  $\int N(\mu, \sigma) dx = 1$ .

A plot of the normal distribution shows that it is bell-shaped. (Hence the ubiquitous term "the bell curve".) The maximum probability is at the mean value  $\mu$ , and the parameter  $\sigma$  determines the "width" of the distribution. If a value  $x$  is away from the mean by a distance  $\sigma$ , then its probability is decreased by a factor of  $e^{-1/2} \approx 0.60$ .

Some terminology: A normal distribution  $N(0,1)$  is called a *standard normal* distribution. Sometimes you will also hear the normal distribution called a *Gaussian* distribution. The two terms can be used interchangeably.

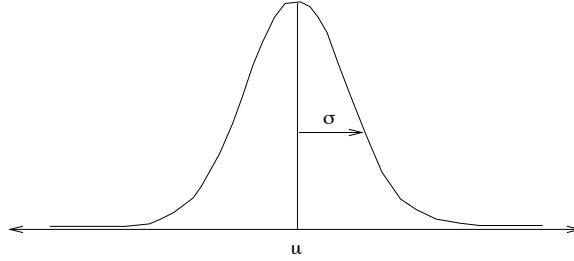


Figure 14:

### 3.2.1 Normalization

Let's use the normalization requirement to determine the constant  $A$ . We require that

$$\int_{-\infty}^{\infty} Ae^{-(x-\mu)^2/2\sigma^2} dx = 1 \quad (48)$$

We wish to solve for  $A$ . There are some tricks to computing this integral, and we'll use them here. To simplify the calculation, let  $y = x - \mu$  and  $dy = dx$ , so that

$$\int_{-\infty}^{\infty} Ae^{-y^2/2\sigma^2} dy = 1 \quad (49)$$

The trick is to write this integral in polar coordinates:

$$\begin{aligned} \int_{-\infty}^{\infty} Ae^{-y^2/2\sigma^2} dy &= A \sqrt{\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx \int_{-\infty}^{\infty} e^{-y^2/2\sigma^2} dy} \\ &= A \sqrt{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2\sigma^2} dx dy} \\ &= A \sqrt{\int_0^{2\pi} d\theta \int_0^{\infty} r e^{-r^2/2\sigma^2} dr} \\ &= A \sqrt{2\pi \left[ -\sigma^2 e^{-r^2/2\sigma^2} \right]_0^{\infty}} \\ &= A \sqrt{2\pi\sigma^2} = 1 \end{aligned} \quad (50)$$

Therefore,

$$A = \frac{1}{\sigma\sqrt{2\pi}} \quad (51)$$

So, the normal distribution is thus:

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (52)$$

This is a very useful piece of information and we recommend you commit it to memory.

### 3.2.2 Properties of the normal distribution

As you may have guessed, the normal distribution is the limit of the binomial distribution as  $N \rightarrow \infty$ , with  $p = 1/2$ . That being the case, let's discover what  $\mu$  and  $\sigma$  correspond to.

**Mean and variance** Recall that, using Bernoulli variables, we found the expectation value  $E(x)$  of the binomial distribution was  $Np$ , and the variance  $E(x^2) - E(x)E(x) = Np(1-p)$ . What is  $E(x)$  for the normal distribution?

We will use the moment-generating function to find  $E(x)$  and  $E(x^2)$ .

$$\begin{aligned} \psi(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2 - 2\sigma^2 tx}{2\sigma^2}} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-\mu-\sigma^2 t)^2}{2\sigma^2}} e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} dx \\ &= e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} \end{aligned} \quad (53)$$

$E(x)$  and  $E(x^2)$  can then be determined by

$$\begin{aligned} E(x) &= \left. \frac{d\psi}{dt} \right|_{t=0} \\ E(x^2) &= \left. \frac{d^2\psi}{dt^2} \right|_{t=0} \end{aligned} \quad (54)$$

$$\begin{aligned} \left. \frac{d\psi}{dt} \right|_{t=0} &= \mu e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} + \sigma^2 t e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} \Big|_{t=0} \\ E(x) &= \mu \end{aligned} \quad (55)$$

Another differentiation yields

$$\begin{aligned} E(x^2) &= \left. \frac{d^2\psi}{dt^2} \right|_{t=0} = \mu^2 + \sigma^2 \\ \text{var}(x) &= E(x^2) - E(x)E(x) \\ \text{var}(x) &= \mu^2 + \sigma^2 - \mu^2 = \sigma^2 \end{aligned} \quad (56)$$

### 3.3 Changes of Variable

Most of the time, statistical methods examine more than the distributions of single variables. We are interested the distributions of *sums* of variables, or distributions of the values of functions of several variables, each of which have their *own* distributions. With this in mind, here we discuss changes of variable.

#### 3.3.1 Finding a coordinate transformation

The basic problem is this: Suppose we know the distribution  $f(x)$  of some quantity  $x$ , and the distribution  $g(y)$  of some related quantity  $y$ , where  $y$  is some function of  $x$ ,  $y = h(x)$ . What is the coordinate transformation  $y = h(x)$ ?

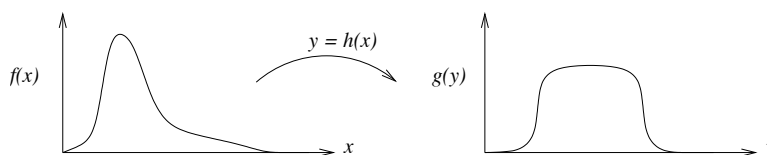


Figure 15:  $y = h(x)$  is the coordinate transformation taking  $x$  to  $y$ .

Suppose the range of both  $x$  and  $y$  run from  $-\infty$  to  $\infty$ . The cumulative probability up to each  $y_0 = h(x_0)$  must be the same:

$$\int_{-\infty}^{x_0} f(x)dx = \int_{-\infty}^{y_0} g(y)dy \quad (57)$$

Given that  $f(x)$  and  $g(y)$  are well-behaved normalizable probability distributions, then according to the above equation, their antiderivatives  $F(x)$  and  $G(y)$  must be the same at  $x_0$  and  $y_0$ :

$$\begin{aligned} F(x_0) &= G(y_0) \\ y_0 &= G^{-1}(F(x_0)) \end{aligned} \quad (58)$$

Thus,  $y = h(x) = G^{-1}(F(x))$  is the coordinate transformation from  $x$  to  $y$ .

#### 3.3.2 Example: generating normal variables

Suppose we need to write a computer program routine that will generate numbers from a normal distribution. Unfortunately, our programming language's built-in `RAND()` function only will generate random numbers from the interval  $[0,1]$ . What is a coordinate transformation that we can apply to  $x \in [0,1]$  to generate random variables from a *normal* distribution  $N(0, \sigma)$ ?

To simplify this problem, we'll only attempt to generate the right side of a normal distribution. (In practice this would be okay, because we could just as easily randomly choose -1 or +1 to multiply the numbers we generated.) Since we define  $g(y)$  over half the usual domain of the normal distribution, it must have twice the usual magnitude to ensure correct normalization:

$$g(y) = \frac{2}{\sigma\sqrt{2\pi}} e^{-y^2/2\sigma^2} \quad (59)$$

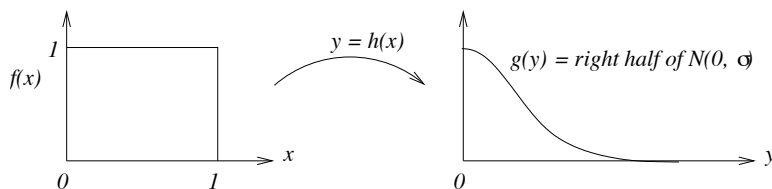


Figure 16:  $y = h(x)$  is the coordinate transformation taking a uniform distribution over  $[0, 1]$  to the right half of a normal distribution with variance  $\sigma$ .

Let  $f(x)$  be our `RAND()` distribution:

$$f(x) = \begin{cases} 1 & \text{if } x \in [0, 1], \\ \text{undefined} & \text{otherwise} \end{cases} \quad (60)$$

and as before, let us equate the cumulative distributions

$$\int_0^{x_0} f(x) dx = \int_0^{y_0} g(y) dy \quad (61)$$

$F(x_0) = x_0$  for  $x_0 \in [0, 1]$ .  $G(y_0)$  is found by the same polar-coordinate trick as before:

$$\begin{aligned} G(y_0) &= \frac{2}{\sigma\sqrt{2\pi}} \int_0^{y_0} e^{-y^2/2\sigma^2} dy \\ &= \frac{2}{\sigma\sqrt{2\pi}} \sqrt{\int_0^{\pi/2} d\theta \int_0^{y_0} r e^{-r^2/2\sigma^2} dr} \\ &= \frac{2}{\sigma\sqrt{2\pi}} \sqrt{\pi/2 [-\sigma^2 e^{-r^2/2\sigma^2}]_0^{y_0}} \\ &= \sqrt{1 - \sigma^2 e^{-y_0^2/2\sigma^2}} \end{aligned} \quad (62)$$

The coordinate transformation is calculated by  $y = h(x) = G^{-1}(F(x))$  to be

$$y = \sqrt{-2\sigma^2 \ln(1 - x^2)} \quad (63)$$

### 3.3.3 Distribution transformations under changes of variables

Let's say that we know the distribution  $f(x)$  of the variable  $x$ , and we have a pre-defined coordinate transformation  $y = h(x)$  to a new variable  $y$ . What is  $g(y)$ , the distribution of  $y$ ?

To find  $g(y)$ , we equate the two cumulative distributions, and use the differentiation *chain rule*:

$$\int g(y) dy = \int f(h^{-1}(y)) \frac{d}{dy} h^{-1}(y) dy \quad (64)$$

Here, of course, we assume that  $h(x)$  is a one-to-one function of  $x$ , so that we can take its inverse. The distribution of  $y$  must then be

$$g(y) = f(h^{-1}(y)) \frac{d}{dy} h^{-1}(y) \quad (65)$$

**Multivariate functions.** Let's say that  $f$  and  $g$  are functions of multiple variables,  $f(x_1, x_2, \dots, x_n)$  and  $g(y_1, y_2, \dots, y_n)$ . If there exists some one-to-one mapping between coordinates  $(y_1, y_2, \dots, y_n) = \mathbf{h}(x_1, x_2, \dots, x_n)$  (such that  $\mathbf{h}$  is invertible), then the chain rule generalizes as follows: Let  $\mathbf{v} = \mathbf{h}^{-1}$ , such that  $(x_1, x_2, \dots, x_n) = \mathbf{v}(y_1, y_2, \dots, y_n)$ . Then

$$g(y_1, y_2, \dots, y_n) = |J| f(\mathbf{v}(y_1, y_2, \dots, y_n)) dy_1 dy_2 \dots dy_n \quad (66)$$

where  $|J|$  is the *Jacobian determinant*, defined as

$$|J| = \begin{vmatrix} \frac{dv_1}{dy_1} & \cdots & \frac{dv_1}{dy_n} \\ \vdots & & \vdots \\ \frac{dv_n}{dy_1} & \cdots & \frac{dv_n}{dy_n} \end{vmatrix} \quad (67)$$

This result is derived using multivariate calculus. It is not important to derive it here, but we show it to simply point out that the chain rule has a generalization for multivariate distributions.

**Variable transformations that aren't one-to-one.** Consider a coordinate transformation that is *not* one-to-one. For example, say  $x_1$  and  $x_2$  are drawn from the interval  $[0,1]$ , and we want to find the distribution of the quantity  $y = h(x_1, x_2) = x_1 + x_2$ . Note that the function  $y = h(x_1, x_2)$  is not invertible, because several combinations of  $x_1$  and  $x_2$  can lead to the same value of  $y$ .

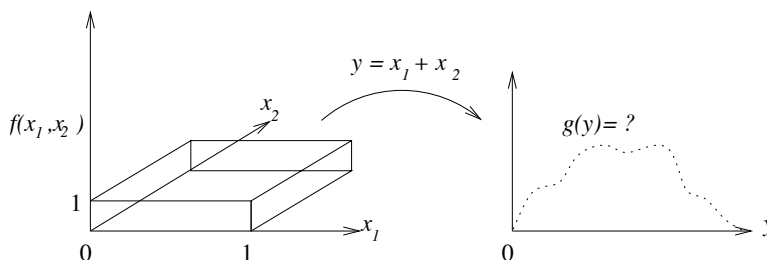


Figure 17: An example of a one-way change of variable transformation  $y = h(x_1, x_2)$ .

We can find the the distribution  $g(y)$  by considering  $x_2 = y - x_1$  and seeing that each chunk of probability  $g(y)dy$  can be calculated by considering the integral of all possible values  $x_2$  that would contribute to this chunk:

$$\int \int f(x_1, y - x_1) dx_1 dy = \int g(y) dy \quad (68)$$

Figure 18: The chunk  $g(y)dy$  is equal to the chunk  $f(x_1, y - x_1)dx_1dy$  taken over all  $x_1$

Let's integrate over  $x_1$  and find an expression for the distribution  $g(y)$ . Since  $x_1$  and  $x_2$  are independent variables,  $f(x_1, x_2) = f(x_1)f(x_2)$ , where  $f(x)$  is our friendly `RAND()` function from the last example.

$$\begin{aligned}
\int_0^{y_0} \int f(x_1, y - x_1) dx_1 dy &= \int_0^{y_0} \int f(x_1) f(y - x_1) dx_1 dy \\
&= \int_0^{y_0} dy \begin{cases} \int_0^y f(x_1) f(y - x_1) dx_1 & \text{if } 0 \leq y < 1, \\ \int_{y-1}^1 f(x_1) f(y - x_1) dx_1 & \text{if } 1 \leq y \leq 2 \end{cases} \\
g(y) &= \begin{cases} y_0 & \text{if } 0 \leq y_0 < 1, \\ 2 - y_0 & \text{if } 1 \leq y_0 \leq 2 \end{cases}
\end{aligned} \tag{69}$$

### 3.4 The Central Limit Theorem

Let  $x_1, x_2, \dots, x_N$  be independent random variables with the same distribution function  $f(x)$ .  $f(x)$  can be any arbitrary distribution function, but with the requirement that  $E(x_i) = \mu$  and  $\text{var}(x_i) = \sigma^2$ .

The *central limit theorem* states that distribution of

$$Z = \frac{\sum_{i=1}^N x_i - N\mu}{\sqrt{N}\sigma} \tag{70}$$

approaches a standard normal distribution in the limit  $N \rightarrow \infty$ .

This powerful statement is the basis for using the normal distribution as a model of variation in measured quantities. It says that quantities determined by many additive effects/errors of measurement

The proof of the central limit theorem is very involved, so we will not present it here. (It involves first showing that if the sequence of moments  $\psi^m$  converges to a particular limit, then the corresponding distributions must also converge; and then goes on to expand the  $\psi^m$ 's by a Taylor series, and show that in the limit of large  $N$  is the moment-generating function of a standard normal distribution.)

### 3.5 The $\chi^2$ Distribution

*Note:* The  $\chi^2$  distribution and the  $t$ -distribution (described in the next section) form the foundational distributions of classical statistics (where it is usually assumed that data is sampled from a Gaussian distribution). In practice, it is usually not necessary to know these (rather complicated) functional forms of these distributions, but instead how to use these distributions when performing statistical tests. We will talk more about this when we get to hypothesis testing.

Suppose we have a set of  $N$  measurements  $m_i$  that we are comparing to expected values from some theoretical distribution  $\hat{m}_i$ . The  $\chi^2$  value is a statistical measure of how close the predictions are, calculated by:

$$\chi^2 = \sum_{i=1}^N \frac{(m_i - \hat{m}_i)^2}{\hat{m}_i} \tag{71}$$

What is the distribution of  $\chi^2$ ? This is a change of variable problem.

If we assume that the  $\chi^2$  value is a variable that is a sum of standard normal variables, then can calculate what the distribution of  $\chi^2$  must be. Let  $x_1, x_2, \dots, x_n$  be standard normal variables, and define

$$Z = x_1^2 + x_2^2 + \dots + x_n^2 = r^2 \quad (72)$$

where  $r$  describes the distance from the origin in an  $N$ -dimensional space.

We want to find the probability distribution for  $Z$ , call it  $F(Z)$ . Since each  $x_i$  comes from a standard normal, we find the coordinate change by:

$$\begin{aligned} \int F(Z)dZ &= \frac{1}{(\sqrt{2\pi})^n} \int e^{-\frac{x_1^2+x_2^2+\dots+x_n^2}{2}} dx_1 dx_2 \dots dx_n \\ &= \frac{1}{(\sqrt{2\pi})^n} \int e^{-\frac{r^2}{2}} \Omega_n r^{n-1} dr \end{aligned} \quad (73)$$

Here we have rewritten the integral in "polar coordinates", but in a high-dimensional space.  $\Omega_n r^{n-1}$  is the hyper-volume of a hyper-spherical shell of width  $dr$  at radius  $r$ . Substituting  $Z$  for  $r^2$ :

$$\int F(Z)dZ = \frac{1}{(\sqrt{2\pi})^n} \int e^{-\frac{Z}{2}} \Omega_n Z^{\frac{n-1}{2}} \frac{dz}{2Z^{1/2}} \quad (74)$$

Let  $C_n$  be the constant value  $C_n = \frac{\Omega_n}{2(\sqrt{2\pi})^n}$ . Then

$$C_n \int e^{-\frac{Z}{2}} Z^{\frac{n}{2}-1} dz = \int F(Z)dZ \quad (75)$$

$$F(Z) = C_n e^{-\frac{Z}{2}} Z^{\frac{n}{2}-1} \quad (76)$$

It turns out that this distribution is a *gamma distribution*, a distribution that is based on the *gamma function*. Here we must interject and introduce these new terms.

**The gamma function.** The gamma function is defined

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad (77)$$

It can be thought of as an extension of factorial function, but for continuous values:  $\Gamma(\alpha + 1) = \alpha!$ . This is because it has the following properties:

Property 1.  $\Gamma(1) = 1$ .

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1 \quad (78)$$

Property 2.  $\Gamma(2) = 1$ .

$$\begin{aligned}\Gamma(2) &= \int_0^{\infty} xe^{-x} dx \\ &= -xe^{-x} \Big|_0^{\infty} + \int_0^{\infty} e^{-x} dx = 1\end{aligned}\tag{79}$$

Property 3.  $\Gamma(\alpha + 2) = (\alpha + 1)\Gamma(\alpha + 1)$ .

$$\begin{aligned}\Gamma(\alpha + 1) &= \int_0^{\infty} x^{(\alpha+1)-1} e^{-x} dx \\ &= \left[ \frac{x^{\alpha+1}}{\alpha+1} \right]_0^{\infty} + \int_0^{\infty} \frac{x^{\alpha+1}}{\alpha+1} e^{-x} dx \\ &= \frac{\Gamma(\alpha + 2)}{\alpha + 1} \\ \Gamma(\alpha + 2) &= (\alpha + 1)\Gamma(\alpha + 1)\end{aligned}\tag{80}$$

**The gamma distribution.** The generalized Gamma distribution is any distribution of the form:

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}\tag{81}$$

Now let's return to our  $\chi^2$ -distribution. We now can see that  $F(z)$  is a Gamma distribution with  $\lambda = 1/2, \alpha = n/2$ :

$$F(Z) = \frac{1}{2^{n/2}\Gamma(n/2)} Z^{\frac{n}{2}-1} e^{-\frac{Z}{2}}\tag{82}$$

This is called a  $\chi^2$ -distribution *with  $n$  degrees of freedom*. Classical statisticians have long computed comprehensive tables listing the probability of obtaining a certain  $\chi^2$  value as a function of  $n$ .

### 3.6 The $t$ -distribution

**Definition.** If  $Z$  is a quantity from a standard normal distribution, and  $U$  is a quantity from  $\chi^2$ -distribution with  $n$  degrees of freedom, and we assume that  $Z$  and  $U$  are independent distributions, then the quantity

$$\frac{Z}{\sqrt{U/n}}\tag{83}$$

has what is called a  $t$ -distribution:

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{(\sqrt{n\pi})\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}\tag{84}$$

Like the other distributions, this result can be obtained directly from a change of variable calculation, but we won't derive it here.

**Application: sample mean and variance** Let  $x_1, x_2, \dots, x_N$  be independent normal random variables with mean  $\mu$  and  $\text{var}(x_i) = \sigma^2$ . Additionally, we define the variables  $\bar{x}$  to be the *sample mean*, and  $s^2$  to be the *sample variance*:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2 \quad (85)$$

In statistics,  $\bar{x}$  and  $s^2$  are often used as estimates of the true mean and true variance of the distribution based on the available data.

The distribution of the quantity

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (86)$$

is the  $t$ -distribution with  $n-1$  degrees of freedom. To see this, let's rewrite this quantity as

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{s^2/\sigma^2}} \quad (87)$$

According the central limit theorem, the top term is a standard normal distribution, and the bottom term has a  $\chi^2$ -distribution with  $n-1$  degrees of freedom, according to its definition.