Modeling and Simulation
Introduction to Probability Theory.

Giampiero Pecelli
Computer Science Department
University of Massachusetts Lowell
Lowell, MA 01854
giam@cs.uml.edu

(978)934-3639
Abstract. This section covers various elementary topics in Probability Theory. It supplements, and overlaps, the material in the text by Law & Kelton.
Chapter 1

An Introduction to Probability Theory.

Before we go on with an attempt at characterizing traffic, we review some definitions and central results from Probability Theory. Although this course does NOT assume extensive exposure to the subject matter, it requires a willingness to "learn-as-you-go". There are a number of perfectly good texts that cover all the necessary material to any depth one cares to reach. One such, which will provide most of the detailed discussion of Probability Theory and Stochastic Processes needed for a detailed understanding of out material, is Ronald W. Wolff: Stochastic Modeling and the Theory of Queues, Prentice-Hall, 1989. Our textbook covers most of the needed material - under the assumption that the student has had some exposure to the subject matter. Another (older and classical) source is W. Feller: An Introduction to Probability Theory and its Applications, Vol. 1, John Wiley, 1957.

1.1 Some definitions.

We start with a set $\Omega$, which we call the sample or probability space, and whose members will be denoted by $\omega$ or some other suitable symbol and will be called outcomes (presumably of some experiment). An event will simply be a subset of $\Omega$, and will be generally denoted by a capital Roman letter. Standard set theoretic notation will be used throughout. We will limit ourselves to examining collections of events, i.e. subsets of the power-set of $\Omega$, and we will use $\mathcal{A}$ to denote such sets of sets ($A \in \mathcal{A}$).

- $\Omega$ is called the certain event.
- The null set, denoted by $\emptyset \subseteq \Omega$, is also called the impossible event.
- If $\Omega$ has finite cardinality, we can let $\mathcal{A}$ be any subset of the power set of $\Omega$, essentially because we can obtain the size of any element of $2^\Omega$ in a completely unambiguous manner - each subset of $\Omega$ has finite cardinality.
- If $\Omega$ is infinite, $\mathcal{A}$ cannot be arbitrary. We must assume that it possesses the following properties:
  - If $A \in \mathcal{A}$ then $A^c = \Omega - A \in \mathcal{A}$.
  - If $A_i \in \mathcal{A}$, $i = 1, 2, \ldots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$. 

3
The first two items can be used to show that countable intersections of elements of $\mathcal{A}$ are also in $\mathcal{A}$.

- The last three properties allow for the definition of measures on $\Omega$, and thus the possibility of replacing summation - which is not (usually) meaningful for uncountably infinite sets - with integration.

**Definition.** A Probability Measure on $\mathcal{A}$ is a function $P$ that assigns a non-negative number to each event in $\mathcal{A}$. Recall that, if $\Omega$ is finite, $2^\Omega$ is also finite and can be $\mathcal{A}$ - i.e., can always have a probability measure on it. We also need some further properties:

**Axiom 1)** $P(A) \geq 0$.

**Axiom 2)** $P(\Omega) = 1$.

**Axiom 3)** If $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$.

Whenever $A \cap B = \emptyset$, we say that the events $A$ and $B$ are mutually exclusive.

A number of elementary results can be obtained without difficulty:

a) From $A \cup A^c = \Omega$, and $A \cap A^c = \emptyset$, along with Axioms 2 and 3, we have $P(A^c) = 1 - P(A)$.

b) From a), we have $P(\emptyset) = 0$.

c) From a) and Axiom 1, we have $P(A) \leq 1$.

d) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

e) $A \subseteq B \Rightarrow P(A) \leq P(B)$.

When $\Omega$ is infinite, we strengthen Axiom 3 to

**Axiom 3’)** Countable Additivity: If $A_1, A_2, \ldots$ are pairwise mutually exclusive - $A_i \cap A_j = \emptyset$ $\forall i \neq j$ - then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (1.1.1)$$

**Claim.** One can show that Countable Additivity is equivalent to the monotone property: Let $\{A_n\}$ be a monotone sequence of events - either $A_1 \subset A_2 \subset \ldots$, or $A_1 \supset A_2 \supset \ldots$ - with $A = \bigcup_{n=1}^{\infty}$ or $A = \bigcap_{n=1}^{\infty}$, respectively. In either case, $\lim_{n \to \infty} P(A_n) = P(A)$.

1.1.1 Conditional Probability: Bayes’ Theorem.

This allows us to compute the probability of an event, *given* that another event has already occurred. Algebraic manipulations allow us to compute some related probabilities.
1.1. SOME DEFINITIONS.

It is easy to see, via a bit of "set counting", that, given two events \( A \) and \( B \),

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

(1.1.2)

is quite meaningful, if we interpret \( P(B|A) \) as "the probability that \( B \) occurs, given that \( A \) has already occurred". We will use this as the definition of Conditional Probability, and we note that the definition is meaningful both in the case of finite \( \Omega \), and, in measure theoretic terms, in case of infinite \( \Omega \).

**Bayes’ Theorem.** Let \( A_i, i = 1, 2, \ldots \) be a sequence of events in \( \mathcal{A} \), and let \( B \in \mathcal{A} \). Then

\[
P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}
\]  

(1.1.3)

**Definition.** When the set \( \{A_i\} \) is both exhaustive - \( \bigcup_i A_i = \Omega \) - and pairwise mutually exclusive, \( \{A_i\} \) is called a partition of \( \Omega \).

Note that, if \( \{A_i\} \) is a partition, then \( B = \bigcup_i (A_i \cap B) \), and, finally,

\[
P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}
\]  

(1.1.4)

where we have used (2) above and Axiom 3'.

**Definition.** We say that event \( B \) is independent of event \( A \) if \( P(B|A) = P(B) \). This assumes that \( P(A) > 0 \). If \( P(B) > 0 \), we obtain immediately that \( P(A|B) = P(A) \), thus giving us a symmetric relationship. Another way of writing this is \( P(A \cap B) = P(A)P(B) \). We can thus talk of independent events.

**Definition.** Let \( \mathcal{C} \) be a class of events. The events of \( \mathcal{C} \) are said to be mutually independent if for every finite subclass \( \{A_1, A_2, \ldots, A_n\} \in \mathcal{C} \),

\[
P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n).
\]  

(1.1.5)

That this is not automatic from the case of two events, can be seen by examining the following problem: suppose a fair coin is tossed twice. \( A \) denotes the event: heads on the first toss; \( B \) denotes the event: heads on the second toss; \( C \) denotes the event: the number of trials on which heads occur is 0 or 2. If the coin is, as assumed, fair, \( P(C|A) = P(C|B) = P(C) = \frac{1}{2} \), thus verifying pairwise independence, while \( P(C|A \cap B) = 1 \neq P(C) \).

1.1.2 Random Variables.

Given a probability triple \( (\Omega, \mathcal{A}, P) \), consisting of a "universal set", a class of events, and a probability measure on the class of events, a random variable \( X \) is a real-valued function \( X : \Omega \to \mathbb{R} \) such that, for every real \( x \),

\[
\{\omega : X(\omega) \leq x\} \in \mathcal{A}.
\]  

(1.1.6)
With this we can define a function \( F : \mathbb{R} \to \mathbb{R} \), by \( F(x) = P(\{\omega : X(\omega) \leq x\}) \). We will use the notation \( X \sim F \) to denote this state of affairs: \( X \) is distributed as \( F \) and \( F \) is the cumulative distribution function of \( X \). We will also use the shorthand \( F(x) = P(X \leq x) \).

A number of properties of \( F \) follow immediately:
For every \( x, 0 \leq F(x) \leq 1 \).
\( F \uparrow x \), i.e. \( F \) is monotone increasing in \( x \).
\( P(X > x) = 1 - F(x) \) is called the tail distribution, or tail of \( F \).

We are now going to examine distributions in more detail.

1.1.3 Discrete Distributions.
\( X \) is a discrete random variable if, for some finite or countable set \( \{x_i\} \subset \mathbb{R} \), \( P(X = x_i) > 0 \) and \( P(X \in \{x_1, x_2, \ldots\}) = 1 \). The cumulative distribution function \( F(x) \) has discontinuities at all the \( x_i \)'s, and is constant between. When the \( x_i \)'s are non-negative integers, we have an integral-valued random variable. We can now look at some examples.

We start with our usual coin tosses, and we let \( 0 < p < 1 \) denote the probability of a head on any one toss. The tosses are assumed mutually independent. Let \( A_i \) denote the event: heads on trial \( i \). The mutual independence implies that the event
\[
A_1 \cap \cdots \cap A_x \cap A_{x+1}^c \cap \cdots \cap A_n^c
\]
has probability \( p^x q^{n-x} \), where \( q = 1 - p \). Counting all the possible orderings we get
\[
P(X = x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, \ldots, n
\]
Another observation is that \( P(X) \) must be proper, i.e., \( \sum_{x=0}^{n} P(X = x) = 1 \). This follows from summing:
\[
\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p + q)^n = 1^n = 1
\]
This distribution is called the Binomial Distribution.

Another useful distribution is the Geometric Distribution. It arises from a slightly different set of events: \( x \) is the number of tosses prior to the occurrence of the first head. Thus \( P(X = x) = q^x p \). Summing this, we have
\[
\sum_{x=0}^{\infty} q^x p = \frac{1}{1 - q} p = 1
\]
A third (and final) useful distribution is the Poisson Distribution, where
\[
P(X = x) = p_x = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \ldots
\]
which see extensive discussion in traffic modeling.
1.1. SOME DEFINITIONS.

1.1.4 Continuous Distributions.

We will deal with three such. What do we mean by a Continuous Distribution? When \( P(X = x) = 0 \forall x, \) \( X \) is continuous, and \( F(x) \) is a continuous function of \( x \). In this case we assume a little more: \( F \) is differentiable, and there exists a density function \( f(x) \geq 0 \) such that \( \int_{-\infty}^{t} f(x) \, dx = F(t) \forall t \in \mathbb{R} \). We also have

\[
f(x) = F'(x) = -\frac{d}{dx} F^c(x) \equiv -\frac{d}{dx} (1 - F(x))
\]

where \( F^c(x) \) is called the tail or tail distribution.

The Normal Distribution. This is given in terms of the density function

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty
\]

where \( \mu \) and \( \sigma \) are real parameters, with \( \sigma > 0 \). We will use the notation \( X \sim N(\mu, \sigma^2) \) to denote that \( X \) is normally distributed.

The Uniform Distribution. This is simply

\[
f(x) = \begin{cases} 
\frac{1}{b - a}, & a \leq x \leq b, \\
0, & \text{otherwise}, 
\end{cases}
\]

where \( b > a \). \( X \) is said to be uniformly distributed on \([a, b]\).

The Exponential Distribution.

\[
f(x) = \begin{cases} 
\mu e^{-\mu x}, & x \geq 0, \\
0, & x < 0,
\end{cases}
\]

where \( \mu > 0 \). We use the notation \( X \sim \exp(\mu) \).

1.1.5 Joint Distributions.

If \( X_1, X_2, \ldots, X_n \) are random variables, the set \( \{ \omega : X_i(\omega) \leq x_i, i = 1, 2, \ldots, n \} \) is the intersection of events in \( \mathcal{A} \), and is thus in \( \mathcal{A} \) - an event. The joint distribution function of the collection \( X_1, X_2, \ldots, X_n \) is defined for all real \( x_1, x_2, \ldots, x_n \) by:

\[
F(x_1, x_2, \ldots, x_n) \equiv P(X_i \leq x_i, i = 1, 2, \ldots, n)
\]

There are obvious implications of monotonicity: \( F \uparrow x_i \) for every \( i \); of forced values: \( F(-\infty, \ldots, -\infty) = 0 \) and \( F(\infty, \ldots, \infty) = 1 \); of the definition of marginal distribution: \( X_i \sim F_i, \) where

\[
P(X_i \leq x_i) = F_i(x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)
\]

Random variables \( X_1, X_2, \ldots, X_n \) are mutually independent if, for all real \( x_1, x_2, \ldots, x_n \),

\[
F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i).
\]
If $X$ and $Y$ are discrete random variables, we can specify their joint distribution by assigning probabilities: $P(X = x, Y = y) = p(x,y)$, while, if they are continuous, we represent the distribution in terms of a joint density function $f(x,y)$, where

$$P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x,y) dx dy.$$ 

To obtain the density function in one variable, we just sum (or integrate) over the other one.

Tail distributions - $P(X > x, Y > y)$ - will be useful, too.

1.1.6 Expectation.

The arithmetic mean of a finite collection of numbers, $a_1, ..., a_n$ is given by $\sum_{i=1}^{n} \frac{a_i}{n}$. In the general case of a discrete random variable $X$, with probability $p_x$, we define the expected value of $X$ as $E(X) = \sum_x x p_x$. If $X$ is a continuous random variable with density $f$, we define $E(X) = \int_{-\infty}^{\infty} x f(x) dx$.

A few of the basic rules are given by:

$$E(cX) = cE(X)$$
$$E(X + Y) = E(X) + E(Y)$$
$$E(u(X)) = \int_{-\infty}^{\infty} u(x) f(x) dx$$

where the last rule can be restated quite easily in the discrete context.

A useful example involving these rules is given by the problem of finding the EXPECTED number of heads in $n$ independent coin tosses. We know that $X$ has the binomial distribution

$$p_x = \binom{n}{x} p^x q^{n-x}$$

so that

$$E(X) = \sum_{x=0}^{n} x p_x = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = np$$

The last equality may not be quite obvious - you may have encountered it in 91.503...

The trick is quite simple: we notice that the expected value of each toss, $E(I_i)$ is given by $E(I_i) = 1p + 0q = p$. Since the tosses are independent, $E(X) = E(\sum_{i=1}^{n} I_i) = \sum_{i=1}^{n} E(I_i) = \sum_{i=1}^{n} p = np$. This is a perfectly good way of proving that the sum has the claimed closed form.

Definition. Higher Moments. The $r^{th}$ moment of $X \sim F$ is defined to be

$$E(X^r) \equiv \int_{-\infty}^{\infty} x^r f(x) dx,$$

with the sum replacing the integral for discrete distributions. The Variance of $X$, denoted by $V(X)$ or by $\sigma^2(X)$, is defined (when $E(X)$ is finite), as $V(X) = E\{[X - E(X)]^2\}$. 
1.1. SOME DEFINITIONS.

1.1.7 Sum of Random Variables.

Correlation and Variance.

The coefficient of correlations is related to the covariance, and we will now introduce these ideas.

The covariance of two random variables $X$ and $Y$ is given in terms of the variance of their sum:

$$V(X + Y) = E((X + Y)^2) - E^2(X + Y)$$
$$= E(X^2) - E^2(X) + E(Y^2) - E^2(Y)$$
$$+ 2E(XY) - 2E(X)E(Y)$$
$$= V(X) + V(Y) + 2Cov(X, Y),$$

and the familiar coefficient of correlation is defined as

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}},$$

where we are assuming the variances to be both positive. The covariance itself is defined as $Cov(X,Y) = E((X - E(X))(Y - E(Y)))$, and can be easily seen to lead to $Cov(X,Y) = E(XY) - E(X)E(Y)$.

In the case of two independent variables we have that

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dx
dy$$
$$= \left[ \int_{-\infty}^{\infty} xg(x)dx \right] \left[ \int_{-\infty}^{\infty} yh(y)dy \right]$$
$$= E(X)E(Y)$$

where the joint density $f(x,y)$ is - by the independence assumption - just the product of the marginal densities $g(x)$ and $h(y)$. Fubini’s Theorem, which gives the splitting of the integral, requires only that the functions in question be measurable (much weaker than continuous) and that the integrals exist.

Since we have just shown that independent variables are uncorrelated, we immediately have the results:

1. If $X$ and $Y$ are uncorrelated (= independent) random variables, then $V(X + Y) = V(X) + V(Y)$.

2. If $U = aX + b$, then $V(U) = a^2V(X)$.

3. if $X_1, \ldots, X_n$ are $n$ pairwise uncorrelated random variables, then

$$V(\sum_{i=1}^{n} c_iX_i) = \sum_{i=1}^{n} c_i^2V(X_i)$$
Exercise: Show that, in general,

\[
V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).
\]

An Application. Let \( X \) be a random variable with binomial distribution

\[
P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots, n,
\]

where \( q = 1 - p \), represented by the sum of independent random variables \( X = \sum_{i=1}^{n} I_i \), where the \( I_i \) terms represent the coin tosses heads \( (I_i = 1) \) and tails \( (I_i = 0) \). Recall that we computed earlier the expected value of \( X \). The argument is essentially identical: since the \( I_i \) are mutually independent, we must have

\[
\begin{align*}
V(I_i) &= E(I_i^2) - E^2(I_i) \\
&= (1^2 \cdot p + 0^2 \cdot q) - (1 \cdot p + 0 \cdot q)^2 \\
&= p - p^2 \\
&= pq,
\end{align*}
\]

individually, and \( V(X) = V(\sum_{i=1}^{n} I_i) = \sum_{i=1}^{n} V(I_i) = n \cdot p \cdot q \).

Distribution of a Sum.

The function that is the resulting distribution of the sum of two random variables is called the convolution of their distributions. You may have run into this term (convolution) in other contexts, e.g., Fourier Transform theory, fast multiplications of polynomials, etc. The ideas are all related, and refer to the same general notion. We will omit examples, for the moment, since the most common (for us) require that we know about Poisson Distributions - which we will introduce later.

We first look at discrete random variables: let \( X \) and \( Y \) be independent with distributions \( P(X = n) = a_n, P(Y = n) = b_n, n = 0, 1, \ldots \). Denote the distribution of \( Z = X + Y \) by \( P(Z = n) = c_n, n = 0, 1, \ldots \). We need to compute \( c_n \) in terms of \( a_n \) and \( b_n \). We observe that

\[
\{Z = n\} = \bigcup_{j=0}^{n} \{X = j, Y = n - j\}, \tag{1.1.11}
\]

a union of mutually exclusive events. Since \( X \) and \( Y \) are independent, we also have

\[
c_n = \sum_{j=0}^{n} P(X = j, Y = n - j) = \sum_{j=0}^{n} a_j b_{n-j}, n = 0, 1, \ldots, \tag{1.1.12}
\]

and \( \{c_n\} \) is called the convolution of \( \{a_n\} \) and \( \{b_n\} \).
1.1. SOME DEFINITIONS.

For the continuous case, let $X$ and $Y$ be independent random variables with density functions $f(x)$ and $g(y)$, and distribution functions $F(x)$ and $G(y)$, respectively. $x, y \in (-\infty, \infty)$. Since we assuming independence, the joint density function is given by $f(x)g(y)$. We first compute the distribution function of $Z = X + Y$:

$$H(t) = P(Z \leq t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-y} f(x)g(y) \, dx \, dy = \int_{-\infty}^{\infty} F(t-y)g(y) \, dy, \quad (1.1.13)$$

where we have used the standard results on double vs. iterated integrals, and the definition of $F(\xi)$ in terms of integrals. An obviously equivalent formulation is given by:

$$H(t) = P(Z \leq t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f(x)g(y) \, dy \, dx = \int_{-\infty}^{\infty} f(x)G(t-x) \, dx. \quad (1.1.14)$$

The density function $h(t)$ can be obtained by computing the derivative of either expression above:

$$h(t) = H'(t) = \frac{d}{dt} \int_{-\infty}^{\infty} \int_{-\infty}^{t-y} f(x)g(y) \, dx \, dy = \int_{-\infty}^{\infty} \frac{\partial}{\partial t} F(t-y))g(y) \, dy$$

$$= \int_{-\infty}^{\infty} f(t-y)g(y) \, dy = \int_{-\infty}^{\infty} f(x)g(t-x) \, dx. \quad (1.1.15)$$

To denote convolution, we use the notation $H = F * G$. It is easy to see that the binary operation of convolution is both commutative and associative.
1.2 Poisson Processes.

Poisson Process: 1)  

The first characterization comes (understandably) from the appendix (p. 336) of [Schwartz 1996]. Consider a small time interval $\Delta t$, and imagine that the discussion following refers to happenings as $\Delta t \to 0$.

i) For arbitrarily small time intervals $\Delta t$, the probability of one arrival during the time interval $\Delta t$ is $\lambda \cdot \Delta t \ll 1$, for some fixed $\lambda > 0$.

ii) The probability of no arrivals in the time interval $\Delta t$ is $1 - \lambda \Delta t$. This simply means that, the probability of more than one arrival in the interval $\Delta t$ is negligible.

Poisson Process: 2)  

Other definitions do not depend on our having a vague notion of smallness. One such is due to W. Feller's book - p. 398. Let $X(t)$ be an integer valued random variable representing the number of arrivals in the time interval $[0, t]$, with $X(0) \geq 0$. Furthermore, one can represent the the increment $X(t+s) - X(0) = [X(t+s) - X(s)] + [X(s) - X(0)]$.

i) The increments $X(s) - X(0)$ and $X(t+s) - X(s)$ are stochastically independent.

ii) The distribution of $X(t+s) - X(s)$ depends only on $t$, i.e., only on the length of the interval and not on its location in $[0, \infty)$.

Let $h_n(t)$ be the probability that $X(t+s) - X(s)$ takes the value $n$, where $n = 1, 2, \ldots$. Since $h_n(t+s) = P(X(t+s) - X(0) = n)$, and $P(X(t+s) - X(0) = n) = P([X(t+s) - X(s)] + [X(s) - X(0)] = n)$, we can apply the results we obtained earlier on the distribution of a sum of independent variables: $h_n(t+s) = \sum_{j=0}^{n} h_j(s) \cdot h_{n-j}(t)$ - a convolution.

We will see later - after we introduce Generating Functions- how this leads to the standard formula for a Poisson Distribution, which we will finally derive in the next subsection.

Poisson Process: 3)  

This is just a slightly more rigorous restatement of our first 'definition', using appropriate little-o notation where needed. $P_n(t)$ denotes the probability that exactly $n$ arrivals (changes, blown lightbulbs, whatever...) will take place during a time interval of length $t \geq 0$. Here are the postulates:

i) The probability that during $(t, t+h)$ an arrival occurs is $\lambda h + o(h)$.

ii) The probability that during $(t, t+h)$ more than one arrival occurs is $o(h)$.

We are going to derive a system of differential equations for $P_n(t), n = 0, 1, \ldots$. Consider two contiguous intervals $(0, t)$ and $(t, t+h)$, where $h$ is small (but otherwise unspecified). If $n \geq 1$, then exactly $n$ arrivals can occur in $(0, t+h)$ in three mutually exclusive ways:

1) there are $n$ arrivals during $(0, t)$ and no arrivals during $(t, t+h)$;

2) there are $n - 1$ arrivals during $(0, t)$ and one arrival during $(t, t+h)$;
3) there are \( x \geq 2 \) arrivals during \((t, t + h)\) and \( n - x \) arrivals during \((0, t)\).

Thus \( P_n(t + h) \) will be the sum of the probabilities of the three disjoint events: the probability of the first is, simply, \( P_n(t) \cdot (1 - \lambda h - o(h)) \); the probability of the second is \( P_{n-1}(t) \cdot (\lambda h + o(h)) \); while the third one has probability \( o(h) \). This give the equation:

\[
P_n(t + h) = P_n(t) (1 - \lambda h) + P_{n-1}(t) \lambda h + o(h).
\]

(1.2.16)

Rearranging and dividing by \( h \), we have:

\[
\frac{P_n(t + h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}.
\]

(1.2.17)

Finally, taking the limit as \( h \to 0 \):

\[
P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t), \text{ for } n \geq 1, t \geq 0.
\]

(1.2.18)

Before continuing, we observe that what appears to be a one-sided limit can be replaced with a two-sided one \((h \leq 0)\), and the derivative is a regular two-sided derivative, with everything is sight being continuous.

We still have the \( n = 0 \) case to take care of. This is easy, and a small amount of thought will lead us to the equation:

\[
P_0'(t) = -\lambda P_0(t), \text{ for } t \geq 0.
\]

(1.2.19)

Since \( P_0(0) = 1 \): the probability that no arrivals occur in the time interval starting at \( t = 0 \) and ending at \( t = 0 \), this last equation comes to us complete with initial condition. It has the unique solution \( P_0(t) = \exp(-\lambda t) \). When we replace this into the equation for \( P_1(t) \), and note that \( P_1(0) = 0 \) (obviously, the probability of any arrivals in a zero-length time interval is 0), we have

\[
P_1'(t) = -\lambda P_1(t) + \lambda e^{-\lambda t}, \text{ for } t \geq 0, \text{ with } P_1(0) = 0.
\]

We now use a standard technique for the solution of non-homogeneous linear differential equations. Since the homogeneous equation (just like 19, with the subscript 1 rather than 0) has the general solution \( Ce^{(-\lambda t)} \), we attempt to solve the non-homogeneous equation (the one with the extra term) by trying the function \( z(t)e^{(-\lambda t)} \) as a candidate for \( P_1(t) \). This leads to the equation

\[
z'(t)e^{(-\lambda t)} - z(t)\lambda e^{(-\lambda t)} = -\lambda z(t)e^{(-\lambda t)} + \lambda e^{(-\lambda t)}.
\]

A trivial simplification leads to \( z'(t) = \lambda \), so that \( z(t) = \lambda t + C_1 \), where \( C_1 \) is an arbitrary constant. Thus \( P_1(t) = (\lambda t + C_1)e^{(-\lambda t)} \). The initial condition \( P_1(0) = 0 \) requires that \( C_1 = 0 \). So, finally, \( P_1(t) = (\lambda t)e^{(-\lambda t)} \).

Using (18), we can show that \( P_n(t) = \frac{(\lambda t)^n}{n!}e^{(-\lambda t)} \) by induction on \( n \), to obtain the desired result: the probability of \( n \) arrivals in \( t \) time units. Thus the probability of at most \( N \) arrivals in \( t \) time units is given by \( \sum_{n=0}^{N} P_n(t) = e^{(-\lambda t)} \sum_{n=0}^{N} \frac{(\lambda t)^n}{n!} \), while that of at least \( N \) arrivals is given by \( \sum_{n=N}^{\infty} P_n(t) = e^{(-\lambda t)} \sum_{n=N}^{\infty} \frac{(\lambda t)^n}{n!} \).
Poisson Process: 4)

The mentioned text by Wolff has another characterization. Essentially, it starts from the probability formula, with some more assumptions, and derives a number of the other properties of the Poisson process.

Let \( \{ \Lambda(t) \} \) be an arrival process; this simply means that \( \Lambda(t) \) gives the number of arrivals in the time interval \([0, t]\), and that \( \Lambda(t) \) is an integer valued non-decreasing function of \( t \) defined on \([0, \infty)\). The quantity \( \Lambda(t + h) - \Lambda(t) \), which is the number of arrivals in \((t, t + h] \), is called the increment in the process between \( t \) and \( t + h \). An arrival process is said to have independent increments if the number of arrivals in disjoint intervals are independent random variables (this is Postulate i) in the second characterization of Poisson processes.

Prop. 1) \( \{ \Lambda(t) \} \) has independent increments.

Prop. 2) \( \exists \lambda > 0 \) such that \( P(\Lambda(t) = j) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \), i.e., \( \{ \Lambda(t) \} \) has a Poisson distribution with mean \( \lambda t \) and arrival rate \( \lambda \).

From the postulates given by Props 1 and 2, we can derive several other properties.

Prop. 3) \( \Lambda(t + h) - \Lambda(t) \sim P(\lambda h) \). This says that increments are stationary random variables - independent of \( t \) - satisfying a Poisson distribution with arrival rate \( \lambda h \). One can derive this result via generating functions - we omit the details for now.

Prop. 4) \( \lim_{h \to 0} P(\Lambda(t + h) - \Lambda(t) \geq 2) / h = \lim_{h \to 0} (1 - P(\Lambda(h) = 0) - P(\Lambda(h) = 1)) / h = \lim_{h \to 0} (1 - e^{-\lambda h}) / h = \lambda e^{-\lambda h} / h \). This property is called orderliness and simply means that arrivals occur one at a time.

Prop. 5) \( \lim_{h \to 0} P(\Lambda(t + h) - \Lambda(t) = 1) / h = \lim_{h \to 0} \lambda e^{-\lambda h} / h = \lambda \). This simply means that the probability of arrival in a short interval is proportional to the interval length.

Poisson Process: 5)

This is the last variant we examine, again from Wolff’s book. We start with \( \{ \Lambda(t) \} \), which simply counts the number of arrivals in the time interval \([0, t]\). Corresponding to this process, there are arrival epochs, \( 0 \leq t_1 \leq t_2, \ldots \), where each \( t_i \) corresponds to the instant of an arrival, and inter-arrival times \( \tau_1 = t_1, \tau_2 = t_2 - t_1, \ldots, \tau_n = t_n - t_{n-1}, \ldots \). From this notation, one can conclude that the event \( \{ \tau_1 > t \} \) occurs \( \iff \{ \Lambda(t) = 0 \} \), since the two events are obviously the same. To obtain the distribution and the density function corresponding to \( \tau_1 \), we observe that \( P(\tau_1 > t) = P(\Lambda(t) = 0) = e^{-\lambda t}, t \geq 0 \), which is the tail of an exponential. The distribution of \( \tau_1 \) is then \( P(\tau_1 \leq t) = 1 - e^{-\lambda t} = F(t) \). Differentiating, we obtain the density function \( f(t) = \lambda e^{-\lambda t}, t \geq 0 \).

We want to show that all \( \tau_j, j \geq 0 \), have the same distribution (and thus the same density function). Conditioning on \( \tau_1 \):

\[
P(\tau_2 > t | \tau_1 = x) = P(\Lambda(t + x) - \Lambda(x) = 0) = e^{-\lambda t},
\]

where we used both independent increments and independence of arrival epochs within \([0, t]\). Notice that the conditional distribution above is independent of the value of \( \tau_1 \),
1.2. POISSON PROCESSES.

which implies that \( \tau_2 \) is independent of \( \tau_1 \) and, clearly, has the same distribution. A similar discussion holds for all \( \tau_j, j > 1 \), and we have that \( \{ \tau_j \} \) is a sequence of independent identically distributed random variables with density function \( f(t) = \lambda e^{(-\lambda t)}, t \geq 0 \).

Our final definition of a Poisson process is of an arrival process \( \{ \Lambda(t) \} \) generated by a sequence \( \{ \tau_j \} \) of interarrival times that have the property just described. We will not show in detail how it all works out.

Conclusion. We have examined several different ways of characterizing Poisson processes. A reason for doing so is that the literature will use whichever characterization is most convenient for the problem being studied, and that being aware of a large context of properties should make our life easier.

1.2.1 Some Moment Computations.

Before we examine an actual model of voice traffic, we look at some of the computed parameters related to Poisson distributions. Let \( P_n(t) \) be the probability of \( n \) arrivals in the time interval \( [0, t] \). Thus \( P_n(t) = \frac{(\lambda t)^n}{n!} e^{(-\lambda t)}, n = 0, 1, \ldots \).

Since the random variable in question is the number of arrivals, we must compute \( E(n, t) \), its expected value, where we have left the dependence on \( t \) explicit.

\[
E(n, t) \equiv \sum_{n=0}^{\infty} n \cdot P_n(t) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{(-\lambda t)}
\]

\[
= e^{(-\lambda t)} \sum_{n=0}^{\infty} \frac{\lambda t}{n!} (\frac{\lambda t}{n})^n = \lambda t e^{-\lambda t} e^{\lambda t} = \lambda.
\]  

(1.2.21)

Remember that the proportionality of the number of arrivals to the length of the time interval was part of the definition of a Poisson process only for small enough intervals. We have now shown that the expected number of arrivals over a time interval is also proportional to the length of the time interval, with the same proportionality constant and with no smallness restriction. Not surprising, but....

What is the average time between arrivals? We saw that the inter-arrival probability density function is \( f(\tau) = \lambda e^{-\lambda \tau} \). Thus the expected inter-arrival time will be given by (we'll use integration by parts):

\[
E(\tau) = \int_0^\infty \tau f(\tau) d\tau = \int_0^\infty \tau \cdot \lambda e^{-\lambda \tau} d\tau = -\tau e^{-\lambda \tau} \bigg|_0^\infty - \int_0^\infty -e^{-\lambda \tau} d\tau
\]

\[
= 0 + \int_0^\infty e^{-\lambda \tau} d\tau = -\frac{1}{\lambda} e^{-\lambda \tau} \bigg|_0^\infty = \frac{1}{\lambda}.
\]  

(1.2.22)

As one would have expected.
In ATM, all packets are exactly the same size, so, although they might arrive at a server with a Poisson arrival distribution, the service distribution - the time to serve a packet - will be uniform. In the IP protocol, since the packets can be of different length, this will not be the case, and it can be assumed that the service distribution is also Poisson, with rate $\mu$ packets/second. The average length of a packet is then $\frac{1}{\mu}$ - we shall derive this below.

Let $r$ be the random variable representing the length of a packet (in, say, seconds). For a Poisson arrival process, the probability density function of $r$ is $f(r) = \mu e^{-\mu r}$. The probability distribution is then

$$P(\text{packet has length } \leq r) = F(r) = \int_0^r \mu e^{-\mu \tau} d\tau = -e^{-\mu \tau} \bigg|_0^r = 1 - e^{-\mu r}$$

The expected length of a packet is given by

$$E(r) = \int_0^\infty r f(r) dr = \int_0^\infty r \mu e^{-\mu r} dr = \frac{1}{\mu}.$$

We can obtain the following result:

Proposition. Packets shorter than average are more likely than packets longer than average.

Proof.

$$P(r \leq \frac{1}{\mu}) = F(\frac{1}{\mu}) = 1 - e^{-\frac{1}{\mu}} = 1 - e^{-1} \approx 0.632121.$$  

On the other hand,  

$$P(r > \frac{1}{\mu}) = 1 - F(\frac{1}{\mu}) = e^{-1} \approx 0.367879.$$  

A very little amount of thought should convince one that this is quite as expected.

Note. Packets are discrete objects, while the theory just developed gives a continuous approximation: expect the possibility of at least some discrepancy between theory and actual behavior. On the other hand, one would expect good agreement between the two when the number of packets transmitted per second is very large: the difference between continuous and discrete approximation decreases with the increase of the rates.

1.2.2 Some More Terminology.

A queueing model with Poisson arrivals, an exponential service time distribution, an infinite buffer and one server is called an $M/M/1$ queue. If the buffer is finite, of size $N$, we speak of an $M/M/1/N$ queue. This terminology was introduced by D. G. Kendall. More precisely, we have the notation $A/B/n/N$, where

1) $A$ is the arrival distribution;

2) $B$ is the service distribution;

3) $n$ is the number of servers;

4) $N$ is the size of the buffer, interpreted as infinite if this last parameter is missing.

$M$ is used to denote the Poisson distribution, $D$ denotes the uniform (constant service time) distribution, while $G$ denotes the general one.
1.3 Packet Voice Modeling.

Speech is not a continuous activity. As it turns out, it consists of short intervals of noise (from 0.4 to 1.2 seconds in duration), alternating with intervals of silence (from 0.6 to 1.8 seconds in duration). This does not take into account the times when we are left speechless by our interlocutor.

We begin by defining two numbers, \( \alpha \) and \( \lambda \), in the following way: \( \frac{1}{\alpha} \) is the length in seconds of the average talk spurt, while \( \frac{1}{\lambda} \) is the length, always in seconds, of the average period of silence. Then \( \alpha \) can be interpreted as the transition rate from talk spurt to silence, while \( \lambda \) is the transition rate from silence into talk spurt. This state of affairs can be represented by the diagram of a two-state birth-death model:

![Diagram](attachment:diagram.png)

**Figure 2: Birth-Death Process**

This particular approach has been used in the telephone world for a very long time (1930s?) to understand what actual traffic a conversation gives rise to so that one can pack more calls into fewer circuits (trunks was the term used by W. Feller in the 1950s). The states (i.e., their durations) are assumed to be exponentially distributed with Poisson parameters within the experimental ranges. How good is this approximation? It appears that talk spurs are reasonably well modeled by Poisson processes, while silence periods are less so. In any case, the use of the models for traffic prediction has been technically and economically successful.

The probability that the speaker is active (that is, in talk spurt) is called the *Speaker Activity Factor*, and is given by \( \frac{1}{\alpha + \frac{1}{\lambda}} = \frac{\lambda}{\alpha + \lambda} \). The probability the speaker is silent is \( \frac{1}{\alpha + \lambda} = \frac{\alpha}{\alpha + \lambda} \). The reciprocal of the Speaker Activity Factor \( \frac{\alpha + \lambda}{\lambda} \) is called the *Time Assigned Speech Interpolation* (TASI for short) advantage. It is easy to see that this value represents the number of conversations that one might be able to (statistically) accomodate in a channel that could support continuous speech.

The case of a single voice source is not very interesting: provide a channel that can support it in talk spurt and you are done. Things get interesting when you try to cheat: if I have \( N \) conversations going between points A and B, do I need to provide \( N \) full channels? Since I would expect some of my speakers to be in silence while others are in talk spurt, I should be able to get my conversations taken care of with only a fraction of the bandwidth.
that I would need to carry them in independent channels - the TASI advantage should be a good indicator of the number of conversations I should be able to support for each real channel - as long as I support the equivalent of many channels in the same connection.

Assume we have \( N \) independent voice sources, and that each voice source generates \( V \) cells/sec while in talk spurt. The \textit{maximum} rate at which cells are generated is \( V \cdot N \), and we propose to use a server whose service capacity is \( V \cdot C \), where \( C < N \) - the opposite inequality being of little interest.

The question we will try to answer is: \textbf{How big must} \( C \) \textbf{be?} Actually, the question is ill posed, since it is not sufficiently specific for an answer. We will start with the \textit{simplest} possible interpretation, in which we worry only about \textit{average} (or expected) values. This will leave the question of what happens during those times when the average is exceeded for a later and more detailed (and technically more complex) look. We will also look at the details for the simple case - the results should be obvious, without going through any of the machinery: this is exactly why we will go through the details of applying the machinery, since, later, the results will be less than obvious and applying the techniques in those cases may be less obvious, too.

Let \( v \) be the random variable that corresponds to number of cells being generated by a user. It takes the values 0 and \( V \). Its expected value is given by:

\[
E(v) = V \cdot \frac{\lambda}{\alpha + \lambda} + 0 \cdot \frac{\alpha}{\alpha + \lambda} = V \cdot \frac{\lambda}{\alpha + \lambda}.
\]

Since we have \( N \) independent identical sources

\[
E(\sum_{i=1}^{N} v) = \sum_{i=1}^{N} E(v) = \sum_{i=1}^{N} V \cdot \frac{\lambda}{\alpha + \lambda} = NV \cdot \frac{\lambda}{\alpha + \lambda},
\]

and this must satisfy the inequality \( NV \frac{\lambda}{\alpha + \lambda} < VC \), or \( N \frac{\lambda}{\alpha + \lambda} < C \), otherwise the queue will become unbounded.

\textbf{Definition.} \( \rho \equiv \frac{\lambda}{\alpha + \lambda} \frac{N}{C} < 1 \), where \( \rho \) is called the \textit{utilization}.

It should be clear that a utilization \textit{less than} 1 must be aimed for; utilizations greater than 1 will lead to unbounded queues (on average) while a utilization of exactly 1 means that half the time (again, on average) the senders are overwhelming the system resources. If the available buffer space is very small, this would mean that packets would have to be dropped with very high probability, while a large buffer might lead to queueing delays that are unacceptable. In any event, we have a first quantitative result that can be used to determine whether a new voice source should be added to the sources in an already existing trunk or not. The result, unfortunately, is not precise enough to be very useful, and we turn to the development of techniques that will allow more useful and more precise estimates.
1.3. PACKET VOICE MODELING.

Since we have \( N \) voice sources, our system could be in any of \( N+1 \) states: \( J_0 \) will denote the state in which none of the sources is speaking; \( J_N \) will denote the state in which all the sources are speaking; \( J_i \) will denote the state in which exactly \( [C] \) sources are speaking (the underload state); and \( J_o \) will denote the state in which exactly \( [C] \) sources are speaking (the overload state). The rate of change of the queue when the system is in state \( i \) is \(-V(C - i)\) cells/second.

The question we can ask now is the following: for \( i \in [0, N] \), what is the probability \( \pi_i \) that the system is in state \( J_i \)?

The answer follows from a simple counting argument: the probability that any one source is in talk spurt is \( \frac{\lambda}{\alpha + \lambda} \), while the probability it is silent is \( 1 - \frac{\lambda}{\alpha + \lambda} = \frac{\alpha}{\alpha + \lambda} \). Since the \( i \) sources can be distributed in any possible way among the \( N \), we have

\[
\pi_i = \binom{N}{i} \left( \frac{\lambda}{\lambda + \alpha} \right)^i \left( \frac{\alpha}{\lambda + \alpha} \right)^{N-i},
\]  

which, after a little manipulation, becomes

\[
\pi_i = \binom{N}{i} \left( \frac{\lambda}{\alpha} \right)^i \left( 1 + \frac{\lambda}{\alpha} \right)^{-N}.
\]

It is trivial to verify that \( \sum_{i=0}^{N} \pi_i = 1 \).

### 1.3.1 Balance Equations.

Another, and to us more useful way of obtaining the result, is provided by a method called the Balance Equation Method. Our text discusses it in the Appendix (p. 339 and following); we will use a slightly different approach as in Feller’s book. Figure 3 shows a diagram corresponding to what is called a Birth-Death Process. It denotes a system with \( N + 1 \) states and the possibility of changing state, but only from the current state to the ones on the left and right, if they exist. Birth would correspond to a jump to the right, while death would correspond to a jump to the left.

![Birth-Death Processes](image)

We have a system with states \( J_0, J_1, \ldots, J_N \). \( N \) could be infinity. Let \( \lambda_n \) be the transition rate from state \( J_n \) to \( J_{n+1} \); let \( \alpha_n \) be the transition rate from \( J_n \) to \( J_{n-1} \).
We can interpret Figure 3 as follows: when in state $J_0$, there are $N$ different speakers who could begin to speak. Since each speaker is responsible for a transition rate from silence into speech of $\lambda$, $N$ such speakers give a transitions rate $N\lambda$ from $J_0$ to $J_1$. Since we can add or subtract only one speaker at a time (this is an assumption which appears reasonable from experience in the case of speech), the transitions can only be as depicted. If we have $i$ speakers speaking, the transition rate up must depend on the number of speakers in silence $(N-i)$, while that down depends on the number of speakers in talk spurt $(i)$.

The axioms for a slightly more general case can be stated as follows:

**Axiom 1**) The system changes only through transitions from states to their NEXT neighbors, in one direction or another.

**Axiom 2**) If at any time $t$ the system is in state $J_n$, the probability that during $(t, t + \Delta t)$ the transition $J_n \rightarrow J_{n+1}$ occurs is $\lambda_n \Delta t + o(\Delta t)$, and the probability of the transition $J_n \rightarrow J_{n-1}$ is $\alpha_n \Delta t + o(\Delta t)$.

**Axiom 3**) The probability that, during $(t,t + \Delta t)$, more than one transition occurs is $o(\Delta t)$.

Let $P_n(t)$ be the probability of finding the system in state $J_n$ at time $t$. We want to compute $P_n(t + \Delta t)$. Note that, at time $t + \Delta t$, the system can be in state $J_n$ only if one of the following conditions is satisfied:

1) The system is in state $J_n$ at time $t$ and no changes occur in $(t,t + \Delta t)$.

2) The system is in state $J_{n-1}$ at time $t$ and a transition to $J_n$ occurs.

3) The system is in state $J_{n+1}$ at time $t$ and a transition to $J_n$ occurs.

4) During $(t,t + \Delta t)$ two or more transitions occur. This event has probability $o(\Delta t)$.

The first three cases are mutually exclusive.

Let’s start from state $J_0$. If we are there at time $t$, at time $t + \Delta t$ we can be either in $J_0$ or in $J_1$. The probability of remaining in state $J_0$ is just $P_0(t)(1 - \lambda_0 \Delta t - o(\Delta t))$, while the probability of reaching $J_0$ from $J_1$ is $P_1(t)(\alpha_1 \Delta t)$. The last possibility is of reaching $J_0$ through more than one transition over the time interval $\Delta t$. Putting these ideas together:

$$P_0(t + \Delta t) = P_0(t)(1 - \lambda_0 \Delta t - o(\Delta t)) + P_1(t) \cdot \alpha_1 \Delta t + o(\Delta t).$$

A similar analysis gives the other boundary condition:

$$P_N(t + \Delta t) = P_N(t)(1 - \lambda_N \Delta t - o(\Delta t)) + P_{N-1}(t) \cdot \lambda_{N-1} \cdot \Delta t + o(\Delta t).$$

The expressions for all the other probabilities depend on movement from both sides, with the formula:

$$P_i(t + \Delta t) = P_i(t)(1 - (\lambda_i + \alpha_i) \Delta t - o(\Delta t)) + P_{i-1}(t) \cdot \lambda_{i-1} \cdot \Delta t + P_{i+1}(t) \cdot \alpha_{i+1} \cdot \Delta t + o(\Delta t).$$

We have a system of $N+1$ equations. Divide all the equations by $\Delta t$, and rearrange, to get, for $i \in [1, N-1]$: 
1.3. PACKET VOICE MODELING.

\[
\begin{align*}
\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} &= -\lambda_0 P_0(t) + \alpha_1 P_1(t) + \frac{\partial P_0(t)}{\partial t} \\
\frac{P_i(t + \Delta t) - P_i(t)}{\Delta t} &= +\lambda_{i-1} P_{i-1}(t) - (\lambda_i + \alpha_i) P_i(t) + \alpha_{i+1} P_{i+1}(t) + \frac{\partial P_i(t)}{\partial t} \\
\frac{P_N(t + \Delta t) - P_N(t)}{\Delta t} &= +\lambda_{N-1} P_{N-1}(t) - \alpha_N P_N(t) + \frac{\partial P_N(t)}{\partial t}
\end{align*}
\]

Taking the limits as \( \Delta t \to 0 \), we end up with a system of \( N + 1 \) Ordinary Differential Equations:

\[
\begin{align*}
P_0'(t) &= -\lambda_0 P_0(t) + \alpha_1 P_1(t) \\
P_i'(t) &= +\lambda_{i-1} P_{i-1}(t) - (\lambda_i + \alpha_i) P_i(t) + \alpha_{i+1} P_{i+1}(t), i \in [1, N-1] \\
P_N'(t) &= +\lambda_{N-1} P_{N-1}(t) - \alpha_N P_N(t).
\end{align*}
\] (1.3.24)

At this moment, we will not attempt to pursue a general solution of the system. We observe that, if our birth-death process reaches a steady-state we must have a constant solution \( P_i'(t) \equiv 0 \), for all \( t \). This implies that the left hand side of the system is always 0, and we have an algebraic system in the \( P_i \)'s - now independent of \( t \). Since the system has \( N + 1 \) equations in \( N + 1 \) unknowns, the determinant of its coefficients must vanish for a nontrivial solution to exist. This can be shown, at least for our special cases. We will start by solving the steady-state algebraic system for \( P_1 \) in terms of \( P_0 \): \( P_1 = \frac{\lambda_1}{\alpha_2} P_0 \). Solving the second equation - the one from the \( P_i'(t) \) term - for \( P_2 \) gives (remember that all the \( P_i'(t) \) terms vanish identically by the steady-state assumption):

\[
P_2 = +\frac{\lambda_1 + \alpha_1}{\alpha_2} P_1 - \frac{\lambda_0}{\alpha_2} P_0 \\
= +\frac{\lambda_1 + \alpha_1}{\alpha_2} \frac{\lambda_0}{\alpha_1} P_0 - \frac{\lambda_0}{\alpha_2} P_0 \\
= \frac{\lambda_0 \lambda_1}{\alpha_1 \alpha_2} P_0.
\] (1.3.25)

Exercise. Prove by mathematical induction that

\[
P_i = \frac{\lambda_0 \ldots \lambda_{i-1}}{\alpha_1 \ldots \alpha_i} P_0 \text{ for all } i > 0.
\]

This gives the solution to the statement on p. 30, formula (3-7), of our text - our denominators are \( \alpha \)'s rather than \( \mu \)'s. The previous discussion sets up the required machinery to make the result understandable.

Our special case, covering \( N \) identical voice sources, gives rise to a slightly simpler system. In particular, we can observe that \( \lambda_i = (N - i) \cdot \lambda \), and \( \alpha_i = N \cdot \alpha \). Renaming \( P_i = \pi_i \), and
\[ \pi = [\pi_0, \pi_1, \ldots, \pi_N], \text{ we have the matrix system:} \]
\[
\pi = \begin{bmatrix}
-\lambda_0 & \lambda_0 & 0 & 0 & \ldots & 0 & 0 \\
\alpha_1 & -(\alpha_1 + \lambda_1) & \lambda_1 & 0 & \ldots & 0 & 0 \\
0 & \alpha_2 & -(\alpha_2 + \lambda_2) & \lambda_2 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & -(\alpha_{N-1} + \lambda_{N-1}) & \lambda_{N-1} \\
0 & 0 & 0 & 0 & \ldots & \alpha_N & -\alpha_N \\
\end{bmatrix} = \mathbf{0}
\]

(1.3.26)

It is easy to see that the columns are linearly dependent - summing them gives the zero column vector - so that the determinant must vanish and the homogeneous linear system has nontrivial solutions.

Denote the previous \((N+1) \times (N+1)\) matrix by \(M\). This matrix is called the **infinitesimal generator** of the Markov process (or Markov chain). It is the coefficient matrix of the original \((N+1)\)-dimensional system of Ordinary Differential Equations that describe the time evolution of the Markov process, expressed as a column vector of probabilities.

We have just seen two different ways of approaching the question: *given* \(N\) users with identical characteristics, *what is the probability that a given number of them will be in talk spurt?* Since we earlier answered the question: *on average, how many such users will overwhelm the bandwidth allocated?*, we have some way of determining whether the available resources have a reasonable probability of doing the job: just compute the probability of being in any one of the states \(J_i, i \in [0, u]\), where \(u\) is the underload index \([C]\). This probability is, obviously \(\sum_{i=0}^{u} \pi_i\). This is not really enough, since the only answer we get, at the moment, could be used to determine what probability of losses we would have if we maintained no queue at the multiplexer. Since voice does allow for some queuing - as long as it does not make packet delivery too late or too uneven - we might be able to do better.

### 1.3.2 Voice Multiplexers with Queues - Fluid Source Modeling.

The question, then, is *how do we estimate queue loss probability?*. We could attempt to answer it by simulation, but simulation would require many experimental runs, and would not alert us to any problem areas that we did not know enough to examine in simulation. Two approaches are discussed in the text, and we will present some more details of one of the two.

The model we will examine is one of a class called **fluid flow** models. The major reason for the name is that we assume all variables to be continuous over the reals - i.e., we dispense with any *discrete* properties - just like in fluid flows. It is obvious that a queue will hold, at any one time, and *integer* number of packets; that rates will be given in bits/second or packets/second; that the number of users will be an integer, and so on. Nevertheless, the transmission rates are so large and a unit change is such a small percentage of the total amount considered, that a continuous approximation should not differ appreciably from the discrete reality. Being able to move to continuous models allows us to use techniques developed over the last several centuries for their quantitative and qualitative analysis. Generally speaking, the models are represented by Ordinary or Partial Differential Equations.
1.3. PACKET VOICE MODELING.

We are going to find out what the size of the queue \( x \) will be, and the associated loss probabilities. The text uses "units of information" rather than packets, and we will follow suit rather than confuse the issue further, since a number of the exercises involve computations based on those units.

Let \( \frac{1}{\alpha} \) be the average length, in seconds, of a talk spurt, and let \( V \) be the rate, in cells/second, at which cells are generated during a talk spurt. If the queue were not served, its size \( x \) would be incremented by \( \frac{V}{\alpha} \) during such a talk spurt. \( \frac{V}{\alpha} \) is called the unit of information, and we will use the variable \( x \) to denote units of information. To convert from cells/second to units-of-information/second (u/s/second) we let \( l \) denote cells, and consider a system with capacity \( VC \) cells/sec. Since there are \( \frac{V}{\alpha} \) cells in a unit of information (ui),

\[
VC \text{ cells/sec } \leftrightarrow \frac{V}{\alpha} (\text{cells/sec})/(\text{cells/ui}) \leftrightarrow \alpha C \text{ ui/sec}
\]

Letting \( l \) denote the number of cells in the buffer and \( x \) denote the number of units of information, we have the conversion formula \( l = x \cdot \frac{V}{\alpha} \). Since we are going to estimate probabilities, we also need the formula \( P(l > i) = P(x > \alpha i/V) \), the probability that the queue contains more than \( i \) cells is equal to the probability that the queue contains more than \( \alpha i/V \) units of information.

What we must do now is compute the probability distribution of \( x \). Once that is done, we can compute the probability of our queue exceeding any preassigned value, whether in terms of units of information or cells.

We define a family of functions \( F_i(t, x), 0 \leq i \leq N, t \geq 0, x \geq 0 \), where \( i \) denotes the number of active sources, \( t \) denotes time, and \( x \) denotes the length of the queue (in units of information). \( F_i(t, x) \) is the cumulative probability distribution that, at time \( t \), and with \( i \) sources active, the queue does not exceed a length of \( x \). For the moment, we put no a priori restriction on the length of \( x \). Queue size would normally be an integer valued quantity. We replace it by a continuous quantity - the general idea is that unit changes are such a small percentage of the quantities managed, that the continuous approximation is appropriate.

We will use balance equations again, and compute the cumulative probability distribution that the system have \( i \) sources active, at time \( t + \Delta t \), and a queue no more than \( x \) in length.
For \( 0 < i < N \), we have

\[
F_i(t + \Delta t, x) = [1 - (N - i)\lambda \Delta t - i\alpha \Delta t] F_i(t, x - i\alpha \Delta t + \alpha C \Delta t) + \\
[N - (i - 1)]\lambda \Delta t F_{i-1}(t, x - (i - 1)\alpha \Delta t + \alpha C \Delta t) + \\
[i + 1]\alpha \Delta t F_{i+1}(t, x - (i + 1)\alpha \Delta t + \alpha C \Delta t) + o(\Delta t).
\]

(1.3.27)

Let us examine the various pieces:

- \([1 - (N - i)\lambda \Delta t - i\alpha \Delta t] \): this is just the probability of staying in state \( i \) (\( i \) sources active), given that you were already in state \( i \). Note that \((N - i)\lambda \Delta t\) is the probability of transition to state \( i + 1 \), while \(i\alpha \Delta t\) is the probability of transition to state \( i - 1 \). The probability of jumping by more than one state is accounted for by the \( o(\Delta t) \) term at the end. Recall that \( \lambda \) and \( \alpha \) are the transition rates from silence into talk spurt and back, respectively. Since in state \( i \) we have \( i \) sources active and \( N - i \) inactive, the transition rate into talk spurt must be \((N - i)\lambda\), while the transition rate into silence must be \( i\alpha \). The transition rate times the time interval provides the transition probability.

- \( F_i(t, x - i\alpha \Delta t + \alpha C \Delta t) \): in order for the queue not to exceed \( x \) at time \( t + \Delta t \), it must not have exceeded some other quantity \( \Delta t \) seconds earlier. The other quantity must be given by the final bound \( x \), from which we subtract the amount added by the \( i \) sources during the time interval \((i\alpha \Delta t)\), and add the amount taken away by the server over the same time interval \((\alpha C \Delta t)\).

- \([N - (i - 1)]\lambda \Delta t \): this is just the transition probability from state \( i - 1 \) to state \( i \). The transition rate is given by the number of inactive sources times the transition rate per source.

- \( F_{i-1}(t, x - (i - 1)\alpha \Delta t + \alpha C \Delta t) \): this gives the state of the queue \( \Delta t \) seconds before the present \((i - 1)\) sources contributing to the increase and the server removing units of information.

- The remaining terms should be now easy to interpret. The only one to be careful about is the \( o(\Delta t) \) one, which accounts for all higher order effects, like the probability of multiple events (transitions) during an interval \( \Delta t \).

The two boundary cases, \( i = 0, N \), give rise to similar equations, except that the equation for \( F_0(t + \Delta t, x) \) contains no contributions from an \( F_{-1} \) term, and the one for \( F_N(t + \Delta t, x) \) contains no contribution from an \( F_{N+1} \) term.

We can rearrange the equations, moving to the left hand side all those terms not containing \( \Delta t \) as a multiplicative factor:

\[
F_i(t + \Delta t, x) - F_i(t, x - (i - C)\alpha \Delta t) = -[(N - i)\lambda + i\alpha]\Delta t F_i(t, x - (i - C)\alpha \Delta t) + \\
[N - (i - 1)]\lambda \Delta t F_{i-1}(t, x - (i - 1 - C)\alpha \Delta t) + \\
[i + 1]\alpha \Delta t F_{i+1}(t, x - (i + 1 - C)\alpha \Delta t).
\]
1.3. PACKET VOICE MODELING.

Dividing through by $\Delta t$ and letting $\Delta t \to 0$, the right hand side becomes:

$$[N - (i - 1)]\lambda F_{i-1}(t, x) - [(N - i)\lambda + i\alpha]F_i(t, x) + [i + 1]\alpha F_{i+1}(t, x)$$

The left hand side can be rewritten as:

$$\frac{F_i(t + \Delta t, x) - F_i(t, x) + F_i(t, x) - F_i(t, x - (i - C)\alpha \Delta t)}{\Delta t}$$

$$= \frac{\partial F_i(t, x)}{\partial t} + \frac{F_i(t, x) - F_i(t, x - (i - C)\alpha \Delta t)}{\Delta t}$$

$$\to \frac{\partial F_i(t, x)}{\partial t} + (i - C)\alpha \frac{\partial F_i(t, x)}{\partial x}, \text{ as } \Delta t \to 0.$$ 

Finally, we have the system of $N + 1$ partial differential equations:

$$\begin{cases}
\frac{\partial F_0(t, x)}{\partial t} = -N\lambda F_0(t, x) + \alpha F_1(t, x) + \alpha C \frac{\partial F_0(t, x)}{\partial x}; \\
\frac{\partial F_i(t, x)}{\partial t} = [N - (i - 1)]\lambda F_{i-1}(t, x) - [(N - i)\lambda + i\alpha]F_i(t, x) + [i + 1]\alpha F_{i+1}(t, x) \\
- (i - C)\alpha \frac{\partial F_i(t, x)}{\partial x}, \text{ for } i = 1, \ldots, N - 1; \\
\frac{\partial F_N(t, x)}{\partial t} = \lambda F_{N-1}(t, x) - N\alpha F_N(t, x) - (N - C)\alpha \frac{\partial F_N(t, x)}{\partial x}.
\end{cases}$$

One could try to look at this analytically or numerically, in an attempt to capture the time dependent evolution of the system. A partial approach starts from the observation that a well-behaved system may, in the long run, approach values that will not change much as time goes on. This is another way of saying that we assume the derivatives to satisfy the conditions

$$\lim_{t \to \infty} \frac{\partial F_i(t, x)}{\partial t} = 0, \text{ for all } i \in [0, N], \text{ and all } x \geq 0,$$

and that all right hand sides are, in the limit, independent of $t$. This gives rise to the system of Ordinary Differential Equations - using $F_i(x)$ to stand for the (now) time-independent cumulative probability distributions:

$$\begin{cases}
-C\alpha F'_0(x) = -N\lambda F_0(x) + \alpha F_1(x); \\
\vdots \\
(i - C)\alpha F'_i(x) = [N - (i - 1)]\lambda F_{i-1}(x) - [(N - i)\lambda + i\alpha]F_i(x) + [i + 1]\alpha F_{i+1}(x); \\
\vdots \\
(N - C)\alpha F'_N(t) = \lambda F_{N-1}(x) - N\alpha F_N(x).
\end{cases}$$

Define $F(x) \equiv [F_0(x), \ldots, F_N(x)]$, a row $(N + 1)$-dimensional vector; $D \equiv \text{diag}[-C\alpha, (1 - C)\alpha, \ldots, (N - C)\alpha]$, an $(N + 1) \cdot (N + 1)$-dimensional diagonal matrix; and let $M$ be the coefficient matrix of the right hand side, which just happens to be the infinitesimal generator of the Markov process discussed some pages back. Then we can re-write the system in more compact form as $F'(x) \cdot D = F(x)M$. If $D$ is invertible - which is equivalent to the constant $C$ not being an integer - we can obtain $F'(x) = F(x)MD^{-1}$, and we are dealing with the elementary theory of linear ODEs.
1.3.3 A Short Refresher Course in Linear ODEs.

We will study the properties of solutions of equations of the form \( y'(t) = y(t)A \), where \( y(t) \) is a function \( y : \mathbb{R} \to \mathbb{R}^n \), for \( n \) a positive integer, and \( y'(t) \) denotes the time derivative of \( y(t) \). It can be shown that such equations do have solutions, i.e., smooth functions that satisfy the equation exist, and that these solutions are unique as soon as we specify some time \( t_0 \) and some point \( y_0 \in \mathbb{R}^n \) and require that \( y(t_0) = y_0 \).

More important questions, for us, are:

- Can we find explicit formulae for the solutions?
- Can we determine the properties that we need to know about?

Solution Methods.

Let’s start with the simplest possible case: a function \( y(t) : \mathbb{R} \to \mathbb{R} \), and a constant \( A \in \mathbb{R} \): \( y'(t) = y(t)A = Ay \). Recall that commutativity would not be preserved in higher dimension. Dividing through by \( y(t) \) we have \( \frac{y'(t)}{y(t)} = A \). So, we are looking for a function which is its own derivative, up to a constant. One can prove that the only such function is the exponential function \( y(t) = K \exp (At) \). The requirement that \( y(t_0) = y_0 \) can be satisfied simply by choosing \( K \) so that \( K \exp (At_0) = y_0 \), or \( K = y_0 \exp (-At_0) \).

A slightly more complex case arises when \( y'(t) = Ay(t) + f(t) \), where \( f(t) \) is some suitable function, usually indicating some external ‘force’ applied to the system. We can obtain a solution of this new system by trying \( y(t) = u(t) \exp (At) \), where \( u(t) \) is an unknown function. Since we want this to satisfy the original equation,

\[
y'(t) = u'(t) \exp (At) + A u(t) \exp (At) = A u(t) \exp (At) + f(t).
\]

Thus \( u'(t) = f(t) \exp (-At) \). An immediate integration gives

\[
u(t) = u(t_0) + \int_{t_0}^{t} f(\tau) \exp (-A\tau) d\tau,
\]

and the general solution is

\[
y(t) = u(t_0) \exp (At) + \exp (At) \int_{t_0}^{t} f(\tau) \exp (-A\tau) d\tau.
\]

The technique generalizes nicely to \( n \)-dimensional systems, as long as we follow certain conventions. In particular, for an \( n \times n \) matrix \( A \), let \( \exp (At) \) denote the infinite series \( \sum_{j=0}^{\infty} \frac{(At)^j}{j!} \), where the sums and products are standard operations on matrices. If \( y(t) \) in an \( n \)-dimensional row vector, then \( y_0 \exp (At) \) will denote that unique solution satisfying \( y(0) = y_0 \). We can go through the same formalism as before to provide formal solutions of non-homogeneous equations - the ones with an extra \( f(t) \) term. Unfortunately, this does not give us enough detail so that we can describe the properties of all these solutions, as the 1-dimensional case allows us to do.
1.3. PACKET VOICE MODELING.

Since any linear combination of any finite number of solutions of the homogeneous equation is a solution, we can try something simpler. Try solutions of the form \( y(t) = Y \exp(\zeta t) \), where \( Y \) is an \( n \)-dimensional constant row vector, and \( \zeta \) is a real or complex number. Allowing \( \zeta \) to be complex will not turn out to be a problem. Replacing in \( y'(t) = y(t)A \) gives \( \zeta Y \exp(\zeta t) = YA \exp(\zeta t) \). We are looking for a pair \( \lambda, Y \), which solves \( Y [A - \zeta I] = 0 \), where \( I \) is the \( n \times n \) identity matrix. This equation has an obvious solution, \( Y = 0 \), the zero row vector, with \( \lambda \) unrestricted, but it is not very useful. It will have non-zero solutions if and only if \( \det(A - \zeta I) = 0 \), where \( \det(X) \) denotes the determinant of \( X \).

It is well known that \( \det(A - \zeta I) \) in an \( n^{th} \) degree polynomial in \( \zeta \). This polynomial has \( n \) zeros, \( \{\zeta_1, \zeta_2, \ldots, \zeta_n\} \), where some of the zeros may be repeated and some may come in complex conjugate pairs (we assume all our coefficients to be real numbers). It is also well-known that, if the zeros (the eigenvalues of the coefficient matrix) are all real and distinct, then distinct non-zero vector solutions \( \{y_1, y_2, \ldots, y_n\} \) exist such that \( y_i (A - \zeta_i I) = 0 \), such that every vector in \( \mathbb{R}^n \) can be written as a unique linear combination of the \( y_i \)'s. If the \( \zeta_i \)s are not all real and distinct things get a little more complicated, but we will not need to concern ourselves with that, at least at the moment.

In analogy with what done in the case of dimension 1, we can try solutions of the form \( y_i(t) = y_i \exp(\zeta_i t) \). We find that

\[
y'_i(t) = \zeta_i y_i \exp(\zeta_i t) = y_i A \exp(\zeta_i t).
\]

and we conclude that every solution of the linear system can be written as a linear combination of the \( \{y_1(t), y_2(t), \ldots, y_n(t)\} \).

1.3.4 Back to the Queueing Model.

We are back to the \((N + 1)\)-dimensional system \( F'(x) = F(x)MD^{-1} \), where \( D \) is an invertible diagonal matrix an \( M \) is the infinitesimal generator of the Markov process, and has the special form we computed. In order for us to proceed, we need to show that - in this special case - the matrix \( MD^{-1} \) has all distinct real eigenvalues. Then every solution of the system must have the form

\[
F(x) = \sum_{i=0}^{N} a_i \Phi_i \exp(\zeta_i x),
\]

where the \( \zeta_i \)s are the distinct eigenvalues of \( MD^{-1} \), the \( \Phi_i \)s are their corresponding eigenvectors, and the \( a_i \)s are appropriate real constants.

Surprising as it may seem, [ANIC 1982] - they talk about 300 baud data rates as normal, so they were not thinking about high speed links as we see them now - managed to prove a number of fairly remarkable properties for the system. We summarize them, leaving the understanding of the details (by reading the original paper) to those so inclined.

- All eigenvalues of \( MD^{-1} \) are real and distinct.
- \( MD^{-1} \) has a zero eigenvalue.
- The number of negative eigenvalues is \( N - |C| \) = the number of overload states.
• The largest negative eigenvalue is given by

\[ r = - \frac{(1 - \rho)(1 + \gamma)}{1 - C/N}, \tag{1.3.28} \]

where \( \gamma = \frac{\lambda}{\alpha} \), and the utilization \( \rho = \frac{\gamma N}{1 + \gamma} < 1 \), with \( C < N \).

The result about the zero eigenvalue follows immediately from our previous discussion, since \( M \) has a zero eigenvalue, but the remaining ones are neither obvious, nor quite trivial.

Since \( F_j(x) \) is the \( j \)th component of the row vector \( F \), letting \( \Phi_i = [\Phi_i,0, \Phi_i,2, \ldots, \Phi_i,N] \), we can represent each of the probability density functions as a sum of the \( j \)th components of the eigensolutions:

\[ 0 \leq F_j(x) = \sum_{i=0}^{N} a_i \Phi_{i,j} \exp(\zeta_i x) \leq 1, \text{ for } j \in [0,N]. \]

What does this require? First of all, the boundedness requirement implies that all the \( a_i \) terms corresponding to positive eigenvalues must be chosen equal to 0; their exponentials are not bounded as \( x \to \infty \), and so they could not represent cumulative probability distributions.

Let’s assume we have found the \( N + 1 \) eigenvalues, and that, without loss of generality, \( \zeta_0 = 0 \). Then \( \Phi_0 = \pi = [\pi_0, \pi_1, \ldots, \pi_N] \), with \( \sum_{i=0}^{N} \pi_i = 1 \), and

\[ F(x) = [F_0(x), F_1(x), \ldots, F_N(x)] = \pi + \sum_{i>0, \zeta_i<0} a_i \Phi_i \exp(\zeta_i x), \]

where \( \pi \) is nothing else but the probability vector corresponding to no restriction on any of the queues: \( \pi_i = F_i(\infty) \). This is the same as the one we computed earlier, providing the probabilities of being in any one of the \( N + 1 \) states.

The text has a worked example with 2 states (one voice source). We shall carry out the computations for a 3 voice system, and four states.
Chapter 2

Generating Functions.

We introduce some further notions from Queueing Theory that have applications to the analysis of switch throughput. We assume a FIFO service discipline, with various types of arrivals (primarily Poisson) and various types of service (primarily Poisson and uniform). The notion we introduce and exploit is that of the Generating Function (or Z-transform). This is really just a special case of a transform - of which Laplace, Fourier, etc... are well known representatives. We start with a definition:

2.1 A Definition.

Let \( n \) be a non-negative-integer valued random variable with probability distribution \( p(n) \) for \( n = 0, \ldots \). The Generating Function of the probability distribution \( p(n) \) is the function

\[
G_n(z) \equiv \sum_{i=0}^{\infty} p(i) z^i. 
\]  

(2.1.1)

Another way of interpreting the expression is that the Generating Function is given by the expectation of the function \( f(n) = z^n \), for \( z \geq 0 \), although \( z \) could even be complex:

\[
E(f) \equiv \sum_{i=0}^{\infty} f(i) p(i) = \sum_{i=0}^{\infty} z^i p(i). 
\]

In case we need to keep track of more than one integer random variable, we will use the notation \( p_n(i) \) to denote the probability that the random variable \( n \) take the value \( i \). In particular, (4.1) would become \( G_n(z) = \sum_{i=0}^{\infty} p_n(i) z^i \).

We can derive some immediate results:

1) \( G_n(1) = \sum_{i=0}^{\infty} 1^i p(i) = \sum_{i=0}^{\infty} p(i) = 1. \)

2) \( G_n(0) = \sum_{i=0}^{\infty} 0^i p(i) = p(0). \)

3) Under some convergence assumptions, we have

\[
\frac{dG_n(z)}{dz} = \sum_{i=0}^{\infty} i z^{i-1} p(i). 
\]
from which we can immediately obtain
\[
\left. \frac{d G_n(z)}{dz} \right|_{z=1} = \sum_{i=0}^{\infty} i \ p(i) = E(n),
\]
the expectation of the random variable.

4) Differentiate once more to obtain
\[
\frac{d^2 G_n(z)}{dz^2} = \frac{d}{dz} \left( \sum_{i=0}^{\infty} i \ z^{i-1} \ p(i) \right) = \sum_{i=0}^{\infty} i \ (i-1) \ z^{i-2} \ p(i).
\]
\[
= \sum_{i=0}^{\infty} i^2 \ z^{i-2} \ p(i) - \sum_{i=0}^{\infty} i \ z^{i-2} \ p(i);
\]
from which we get
\[
\left. \frac{d^2 G_n(z)}{dz^2} \right|_{z=1} = E(n^2) - E(n) = E(n^2) - E^2(n) - E(n) (1 - E(n)) \quad \text{(2.1.2)}
\]
\[
= \sigma^2 - E(n) (1 - E(n)),
\]
providing us with an easy method to compute the variance of the distribution.

5) The higher moments can be computed in a similar fashion.

2.2 Some Simple Examples.

2.2.1 The Poisson Distribution.

Let \( \lambda \) denote the arrival rate - in cells/sec, or whatever units are appropriate - and let \( p_\lambda(i) \equiv \text{the probability of exactly } n \text{ arrivals in a unit time interval.} \) Although \( n \) is the random variable, each distinct choice of \( \lambda \) identifies a unique Poisson Distribution, since, for Poisson distributed arrivals we have:
\[
p_\lambda(i) = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \ldots
\]
We will thus use \( \lambda \), rather than \( n \), to identify the corresponding generating function. In this way:
\[
G_\lambda(z) = \sum_{i=0}^{\infty} z^i \frac{\lambda^i e^{-\lambda}}{i!}.
\]
It is now trivial to see:
\[
G_\lambda(z) = \sum_{i=0}^{\infty} z^i \frac{\lambda^i e^{-\lambda}}{i!} = \sum_{i=0}^{\infty} \frac{z^i \lambda^i e^{-\lambda}}{i!}
\]
\[
= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(z \lambda)^i}{i!} = e^{-\lambda} e^{\lambda z} = e^{-\lambda (1 - z)}
\]
2.2. SOME SIMPLE EXAMPLES.

Differentiate with respect to $z$, and evaluate at $z = 1$ to obtain

$$\frac{d G_\lambda(z)}{dz} \bigg|_{z = 1} = \lambda = E_\lambda(n),$$

and

$$\frac{d^2 G_\lambda(z)}{dz^2} \bigg|_{z = 1} = \lambda^2 = \sigma^2_\lambda - E_\lambda(n) (1 - E_\lambda(n)),$$

from which it is trivial to see that $\sigma^2_\lambda = \lambda$.

2.2.2 The Geometric Distribution.

Let $n$ be a non-negative-integer-valued random variable denoting number of trials. Let $p(i)$ denote the probability of some event occurring exactly once in exactly $n$ trials. Let $p$ denote the probability of the event occurring in one trial, and $q = 1 - p$ the probability of the event not occurring. Then, assuming the trials are independent, $p(i) = p q^{i-1}$, the probability of exactly one success and $i - 1$ failures in $i$ trials. The generating function is

$$G_n(z) = \sum_{i=1}^{\infty} z^i p(i) = \sum_{i=1}^{\infty} z^i p q^{i-1} = z p \sum_{i=0}^{\infty} (z q)^i = \frac{z p}{1 - z q}.$$  

(Why does the sum start at 1 instead of 0?)

It is easy to show, by differentiating and evaluating at $z = 1$, that

1) $E(n) = \frac{d G_n(z)}{dz} \bigg|_{z = 1} = \frac{1}{p}$
2) $\sigma^2_n = \frac{d^2 G_n(z)}{dz^2} \bigg|_{z = 1} + E(n) (1 - E(n)) = \frac{q}{p^2}$.

2.2.3 The Bernoulli Distribution.

The random variable $n$ can take only the values 0 and 1. Then $p(0) = q$ (the probability of no occurrence) and $p(1) = 1 - q = p$ (the probability of occurrence. $n$ can thus denote a source being on or off; a cell arriving or not arriving at a switch during a "cell interval", etc. The generating function is $G_n(z) = z^0 p(0) + z p(1) = q + z p$. It follows immediately that $E(n) = p$ and $\sigma^2_n = 0 + E(n) (1 - E(n)) = p (1 - p) = p q$.

2.2.4 The Binomial Distribution.

Let $N$ be a non-negative integer. Let $n$ be a non-negative-integer valued random variable. Let its probability distribution be given by the formula $p(i) = \binom{N}{i} p^i q^{N-i}$ for $0 \leq i \leq N$ and $p(i) = 0$ for all $i > N$. We also assume that $0 \leq p \leq 1$, and $q = 1 - p$. An interpretation for $n$ is the number of independent sources (out of $N$) that are in the "on" state simultaneously, where $p$ denotes the probability that one source is in the "on" state, and we assume the sources to be identical. The definition of a generating function, plus some simple differentiations and evaluations, using (4.2), give
1) \( G_n(z) = \sum_{i=1}^{N} \binom{N}{i} z^i p^i q^{N-i} = (z p + q)^N; \)

2) \( E(n) = N p \)

3) \( \sigma_n^2 = N p (1 - p). \)

2.2.5 The Sum-Product Property.

If the Bernoulli distribution corresponds to the behavior of one source and the Binomial distribution provides a model for the behavior of \( N \) independent but identical sources, we see that the random variable denoting the number of sources in an "on" state is obtained as the sum of the random variables corresponding to each source. If we refer to the short study of sums of random variables carried out in Sec. 9.8, which led us to conclude that the probability distribution of a sum was given by the convolution of the distributions (convolution is a kind of product) and we observe that, in this special case, the generating function of the sum is given by the \textit{product} of the generating functions, we should be able to conjecture that this is a general result.

**Theorem.** The generating function of a sum of independent random variables is given by the product of the generating functions.

**Proof.** Let \( n_1, \ldots, n_N \) be \( N \) independent random variables. Let \( y = \sum_{i=0}^{N} n_i \) be the random variable corresponding to their sum. Since exponentiation is continuous, the expressions \( z^{n_1}, \ldots, z^{n_N} \) are also independent random variables, at least for the values of \( z \) of interest to us. We have

\[
z^y = z^{\sum_{i=0}^{N} n_i} = \prod_{i=0}^{N} z^{n_i}.
\]

We recall the earlier observation that \( G_n(z) = E(z^n) \), to obtain:

\[
G_y(z) = E(z^y) = E(z^{\sum_{i=0}^{N} n_i}) = E(\prod_{i=0}^{N} z^{n_i}) = \prod_{i=0}^{N} E(z^{n_i}) = \prod_{i=0}^{N} G_{n_i}(z).
\]

QED. (more or less)

**Corollary.** Let \( X \) and \( Y \) be independent Poisson distributed random variables with arrival rates \( \lambda \) and \( \mu \), respectively. Let \( Z = X + Y \). Then \( Z \) is a Poisson distributed random variable with arrival rate \( \lambda + \mu \).

**Proof.** By the previous Theorem,

\[
G_Z(z) = G_X(z) G_Y(z) = e^{\lambda (z-1)} e^{\mu (z-1)} = e^{(\lambda + \mu) (z-1)}.
\]

The generating function of the sum has the desired form.

**Corollary.** Assume we have \( N \) independent, identically distributed random variables with generating function \( G_a(z) \). Their sum \( y \) has the generating function \( G_y(z) = [G_a(z)]^N \).
2.3. AN APPLICATION.

The proof is omitted. This shows that the relationship between the Binomial and Bernoulli distributions is a special case of a general relationship.

2.3 An Application.

We attempt to compute the queue occupancy statistics for a configuration as in Fig. 17.1. Cells arrive at a service point with arrival statistics given by a random variable \( a \), they are served in uniform time - one cell every \( 1/C_L \) seconds - and the excess cells are queued. Let \( n \) denote the random variable representing queue occupancy. What we want is to derive the probability distribution of \( n \), given the probability distribution of \( a \).

![Figure 3.1: Arrivals and Queues.](image)

We will take a hint from the discussion on Admission Control in Sec. 16.5. In particular, we will consider the state of the queue - and thus the value of \( n \) - at an instant just before a cell is to be transmitted. No other events are allowed to take place between this observation and the next cell transmittal time (there may be no actual cell awaiting transmittal). The configuration is presented diagrammatically in Fig. 17.2 below, where \( k \) denotes time in unit increments.

![Figure 3.2: Arrivals and Queues: Timing Considerations.](image)

It is easy to show that the following relation must hold:

\[
n_{k+1} = (n_k - 1)^+ + a_{k+1},
\]

where \( r^+ = r \), if \( r \geq 0 \), and = 0 otherwise.

Let \( x \) denote the random variable \((n - 1)^+\). We can thus write \( n = x + a \): \( n \) must be the sum of the other two random variables. We make the further assumption that \( a \) and \( x \) are independent, apply the Sum-Product Theorem, and conclude that

\[
G_n(z) = G_x(z) G_a(z).
\]
Let $p_n(i), p_x(i), p_a(i), i = 0, 1, 2, \ldots$, be the probability distributions of the three random variables. Since the random variable $x$ is defined in terms of the variable $n$, it should be possible to rewrite the generating function of $x$ in terms of that of $n$. We observe, since $x$ takes the value 0 when $n$ takes the values 0 and 1, and takes the value $i - 1$ when $n$ takes the value $i$ for $i > 1$:

$$G_x(z) = \sum_{i=0}^{\infty} z^i p_x(i)$$

$$= z^0 p_x(0) + z^1 p_x(1) + z^2 p_x(2) + z^3 p_x(3) + \ldots$$

$$= z^0 (p_n(0) + p_n(1)) + z^1 p_n(2) + z^2 p_n(3) + z^3 p_n(4) + \ldots$$

$$= p_n(0) + \frac{G_n(z) - p_n(0)}{z}$$

Replacing $G_x(z)$ in the previous equation, and solving for $G_n(z)$, we have

$$G_n(z) = \frac{p_n(0) (z - 1) G_a(z)}{z - G_a(z)}.$$  \hfill (2.3.3)

The only unknown element is $p_n(0)$, the probability that the queue is empty. We can compute this value by using the well-known fact the $G_X(1) = 1$, for any random variable $X$. Unfortunately, just replacing $z$ by 1 in (6.3) leaves us with an indeterminate form. Differentiate both numerator and denominator of the fraction with respect to $z$, take the limit as $z \to 1$, and apply L’Hospital’s Theorem, to get

$$G_n(1) = 1 = \frac{p_n(0)}{1 - G_a'(1)},$$

which gives $p_n(0) = 1 - E(a)$. Finally,

$$G_n(z) = \frac{(1 - E(a)) (z - 1) G_a(z)}{z - G_a(z)}.$$  \hfill (2.3.4)

**Some Comments.** i) The only assumption made was that the queue service discipline was uniform. No assumptions were made about the arrival discipline, and thus the random variable $a$. Given any arrival statistics, we can now compute the queue statistics. ii) If we know the queue statistics, we can find the arrival statistics, since solving for $G_a(z)$ instead of $G_n(z)$ gives

$$G_a(z) = \frac{z G_n(z)}{(z - 1) p_n(0) + G_n(z)}.$$
Chapter 3

Voice Traffic.

We approach the problem of estimating required queue sizes to meet QoS requirements in a slightly different way. This technique will see some use in the later attempts at characterization of Video Sources, and, although (apparently) not as accurate a model of voice traffic as the previous fluid-flow model, it will provide a reasonable line of attack in the other domain. We will not pursue it in full generality at this point, just introduce it and then move to an application in a special case.

Consider a finite-state stochastic process with \( n \) states, 1, 2, \ldots, \( n \). In state \( i \) the stochastic process is representable via a Poisson Distribution with arrival rate \( \lambda_i \). Let \( \mu_{ij} \) denote the (non-negative) transition rate from state \( i \) to state \( j \).

In order to derive the governing equation, we assume a continuous time process, with \( P_i(t) \) denoting the probability that the process is in state \( i \) at time \( t \).

\[
P_i(t + \Delta t) = (\text{prob. that process does NOT leave state } i) \cdot P_i(t) \\
+ (\text{prob. that process moves INTO state } i \text{ from all other states}) + o(\Delta t) \\
= (1 - \sum_{j \neq i} \mu_{ij} \Delta t) P_i(t) + \sum_{j \neq i} \mu_{ji} \Delta t P_j(t) + o(\Delta t).
\]

Moving the \( P_i(t) \) term to the left hand side, dividing by \( \Delta t \), and letting \( \Delta t \to 0 \), we obtain

\[
P_i'(t) = -\sum_{j \neq i} \mu_{ij} P_i(t) + \sum_{j \neq i} \mu_{ji} P_j(t) \\
= \mu_{ii} P_i(t) - \sum_{j \neq i} \mu_{ji} P_j(t), \text{ for } i = 1, \ldots, n.
\]

The assumption that the system has reached steady state implies that \( P_i'(t) = 0, \forall t \geq 0 \), thus implying that we should look for \( P_i \) values independent of \( t \). We can use the notation \( \mathbf{P}(t) \equiv [P_1(t), \ldots, P_n(t)] = \pi = [\pi_1, \ldots, \pi_n] \), to indicate, in a compact manner, that we are trying to solve a matrix equation \( \pi \mathbf{M} = \mathbf{0} \), where \( \mathbf{M} \equiv [\mu_{ij}]_{i=1,\ldots,n; j=1,\ldots,n} \). It is trivial to observe that \( \det(\mathbf{M}) = 0 \), so that a nontrivial solution \( \pi \), whose components add to 1, exists.
Notice that we have made no specific use of the arrival rates $\lambda_i$ associated with state $i$ of the process. We can re-index each state $i$, say $s_i$, as a sequence of states indexed by the queue-length $k$: we now have states $s_{ik}: k = 0, 1, 2, \ldots$, where $k$ could either increase without bound, or be bounded by some positive value $B$. The state $s_{ik}$ corresponds to the statement (in the context of voice source multiplexing): $i$ sources are in talk spurt and the queue length is exactly $k$.

The example taken up in [Schwartz 96] makes no use of transitions beyond the Birth-Death process type, but uses the transmission rates to analyze queue sizes, while introducing a new set of solution methods. We start with exactly the same configuration as in Sect. 2.3.1:

![Birth-Death Processes](image)

*Figure 3.3: Birth-Death Processes.*

At this point, we assume that the output from the process is served by a statistical multiplexer. We also assume that the packet length, rather than being uniform, will be exponentially distributed with average cell length $1/\nu$ seconds - or average service rate of $\nu$ packets/second.

We have some easy results: with all the symbols we have defined, we must have

$$N \beta \frac{\lambda}{\alpha + \lambda} < \nu,$$

which simply says that the expected number of cells generated per second must not exceed the multiplexer capacity. In terms of the utilization $\rho$, we must have

$$\rho \equiv \frac{N}{\beta \cdot \nu} \frac{\lambda}{\alpha + \lambda} < 1.$$

### 3.1 The Generation of Voice Traffic.

All the mathematical modeling has to be sanity checked - not because the mathematics could be wrong (it could, but because of somebody's mistake rather than some intrinsic reason), but because the mathematical model may have nothing to do with the reality. One of the ways to sanity check it is to run real experiments with real equipment and real-time data: this is rather expensive and very time-consuming. Another way is to run simulations. We need to construct a simulator for the features we want to examine, and the simulator should be 'obviously' appropriate to the questions being asked (this is a problem, but the simulator is usually intuitively much closer to reality than an analytical model). Once we
have the simulator, we need to feed it data. In this case, we need to feed it packets in ways that are consistent with real voice, without using a human to generate the voice. How do we do this?

What do we know about voice? That it is characterized by two numbers: $1/\alpha$ is the average amount of time in talk spurt, while $1/\lambda$ is the average amount of time in silence. We also know that the actual lengths of talk spurt intervals are reasonably well represented by a Poisson distribution, while those of silence are less well represented by a Poisson distribution. At a first approximation, we will assume both talk spurt and silence intervals satisfy Poisson distributions, with means $1/\alpha$ and $1/\lambda$, respectively. A source in talk spurt generates ATM cells at a uniform rate of about 167/sec.

We can probably assume all our sources start out in silence, and begin talking at different times. Let’s concentrate on a single source, say $S$, so that $S$ is silent until time $t_0 \geq 0$. When time $t_0$ arrives the simulator must decide how long the talk spurt for source $S$ will be. We can use the distribution of interarrival times to provide us with the lengths of the talk spurt intervals: $F(t) = \int_0^t \mu e^{-\mu t'} \, dt'$ is the cumulative probability distribution, that is the probability that the interarrival time be not greater than $t$ (see Section 10.3 for some relevant material). The probability density function is just $f(t) = \mu e^{-\mu t}$, and the expected value of the random variable can be easily seen to be $1/\mu$. So, if $1/\alpha$ is the length of the average talk spurt, all we need to do is replace $\mu$ by $\alpha$ in the discussion.

Assume now that we have a random number generator that will generate values $x$ in $[0, 1]$ with uniform probability. Since $0 \leq F(t) < 1$, and $\lim_{t \to \infty} F(t) = 1$, what we do is solve the equation $x = F(t)$ for $t$. More precisely, a bit of elementary calculus gives $t = \frac{1}{\mu} \ln (1 - x)$ for the duration of the action. As before (p. 53), intervals of duration shorter than average will be more frequent than intervals of duration longer than average.

In general, if we have a continuous cumulative probability distribution $F(t)$, solving $x = F(t)$ for $t$ will provide the desired value of the random variable, starting only from a random number generator. If the distribution is discrete $F(n), n = 0, 1, \ldots$, all we need to do is find the interval $[n-1, n]$ such that $x \in [F(n-1), F(n))$. Then $n$ is the desired integral value of the random variable.

As a final example, if a random variable is normally distributed with mean $\mu$ and variance $\sigma^2$, its probability density function is given by

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

Its cumulative distribution function is given by

$$F(t) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^t e^{-(\tau-\mu)^2/2\sigma^2} \, d\tau.$$  

Using the random number generator to generate a normally distributed continuous random variable involves solving $x = F(t)$ for $t$. This could be done via a binary search approximation scheme that would involve a precomputed table - coupled, possibly, with an interpolation method.
A property of the Normal Distribution makes the computations relatively easy: if $X$ is a normally distributed random variable with mean $\mu$ and variance $\sigma^2$, then the random variable
\[ Z = \frac{X - \mu}{\sigma} \]
is normally distributed, with mean 0 and variance 1. One needs to precompute cumulative distribution values for the canonical normal distribution, and one can then convert from one to the other with little trouble and a very small amount of computation.

Continuing with the simulation, at time $t_0$ the source $S$ generates a random number $x_1$, solves the equation $t_1 = \frac{1}{\alpha} \ln (1 - x_1)$, and sets a new time of $t_0 + t_1$. During this time period it must generate ATM cells at the appropriate rate - one every $167^{th}$ of a second. A soon as the clock has exceeded $t_0 + t_1$, the source runs the random number generator again to get a number $x_2$, solves the equation $t_2 = -\frac{1}{\lambda} \ln (1 - x_2)$, and stops generating cells until time $t_0 + t_1 + t_2$. And so on, alternating, until the end of the simulation.

The situation with multiple sources may also involve turning a source on and turning it off, which is different from the talk spurt/silence cycle. You would need some information about the average number of conversations started over a given time interval and the average length of a conversation: you would need to use the statistics to decide if an attempt at adding a new conversation is started, and the information you have about losses to decide whether you can afford to accept it. According to some data, telephone conversations last approximately 3.5 minutes, and their length is (approximately) Poisson Distributed. Thus one could set the switching of sources off and on according to such a probability distribution. The rate at which they start is clearly based on the overall population we are trying to serve, so one could set the rate of startup to be whatever one wants - the problem then becomes one of deciding when an attempt to start a phone call should succeed.

Recall that, in audio, the acceptable length of the queues involved has to be small (no more than a few hundredths of a second total over all the queues encountered end-to-end), so your simulation must bound the queues - dropping packets when there is no space in the queue - and keeping track of the percentage of packets dropped vs. packets received. This latter fraction will give the empirical loss probability.
Chapter 4

Some Test Distributions.

4.1 Generalities.


Most of the probability distributions discussed up to this point are useful to represent various processes. We just saw an application to the generation of voice traffic. One of the problems one must solve is the determination of the distribution parameters of real data so that appropriate synthetic data can be generated.

The discussion about Generating Functions lets us conclude that, if we could compute all the moments of the real data set, we could also retrieve the complete probability distribution. Unfortunately, the real data set is just a (maybe) random subset of the set of all possible data values. How can we possibly determine that we have enough to be able to claim anything?

In order to be able test that something is (plausibly) a desired first or second moment, we need to know something about the distribution of empirical moment values obtained from a sequence of tests. In this direction, we state two useful (essential!) results.

**Theorem: The Law of Large Numbers.** Let \( \{X_k\} \) be a sequence of mutually independent random variables with a common distribution. If the expectation \( \mu = \mathbb{E}(X_k) \) exists, then for every \( \epsilon > 0 \), as \( n \to \infty \),

\[
P\left\{ \frac{X_1 + \cdots + X_n}{n} - \mu > \epsilon \right\} \to 0;
\]

i.e., the probability that the average \( S_n/n \) will differ from the expectation by less than an arbitrarily prescribed \( \epsilon \) tends to 1. Or: the sample mean is a legitimate approximation for the expected value of the population, if the sample is large enough. [Feller, p. 228]

**Theorem: The Central Limit Theorem.** Let \( \{X_k\} \) be a sequence of mutually independent random variables with a common distribution. Suppose that \( \mu = \mathbb{E}(X_k) \) and
CHAPTER 4. SOME TEST DISTRIBUTIONS.

\[ \sigma = \text{Var}(X_k) \text{ exist and let } S_n = X_1 + \cdots + X_n. \] Then, for every fixed \( \beta > 0 \),

\[ P \left\{ \frac{S_n - n \mu}{\sigma \sqrt{n}} < \beta \right\} \to \Phi(\beta), \]

where \( \Phi(\beta) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta} e^{-(y^2)/2} \, dy \), the canonical normal distribution. [Feller, p. 229] i.e., the sample mean tends (for large enough samples) to be normally distributed with mean equal to the population mean and standard deviation equal to the population standard deviation divided by the square root of the sample size. Note that this does NOT depend on the probability distribution of the original population - beyond satisfying some reasonable conditions.

Both theorems can be stated and proven in much more general forms, but we will limit ourselves to these simpler ones: they are easy to understand and interpret. We will now give an example application for the Central Limit Theorem, taken from [Feller, p. 230].

Suppose that in a population of \( N \) families there are \( N_k \) families with exactly \( k \) children \( (k = 0, 1, \ldots; \sum N_k = N) \). For a family chosen at random, the number of children is a random variable which assumes the value \( \nu \) with probability \( p_\nu = N_\nu/N \). A sample of size \( n \) with replacement represents \( n \) independent random variables \( X_1, \ldots, X_n \) each with the same distribution. \( S_n/n \) is the sample average.

The Law of Large Numbers tells us that for sufficiently large random samples the sample average is likely to be near \( \mu = \sum \nu p_\nu = \sum \nu N_\nu/N \), the population average. The Central Limit Theorem allows us to estimate the probable magnitude of the discrepancy and to determine the sample size for reliable estimates, even though, in practice, both \( \mu \) and \( \sigma^2 \) are unknown. Let’s say that we want to determine what sample size will give us a probability of 0.99 that the sample average \( S_n/n \) differs from the unknown population mean \( \mu \) by less than \( \frac{1}{10} \). The sample size should be such that

\[ P \left\{ \left| \frac{S_n - n \mu}{n} \right| < \frac{1}{10} \right\} \geq 0.99. \]

We can manipulate this into the equivalent

\[ P \left\{ \left| \frac{S_n - n \mu}{\sqrt{n}} \right| < \frac{1/2}{\sigma} \right\} \geq 0.99. \]

The next step is to notice that, using the notation of the Central Limit Theorem, \( \beta = n^{1/2}/(\sigma 10) \), and that we want the solution to \( \Phi(x) - \Phi(-x) = 0.99 \) (the left tail of the canonical normal distribution is also rejected). A numerical solution of the transcendental equation - or a table lookup - gives that \( x \approx 2.57 \), and hence \( n \) should satisfy \( n^{1/2}/(10 \sigma) \geq 2.57 \), or \( n \geq 660 \sigma^2 \). We still don’t know what \( \sigma \) is, but we can probably reach a conservative estimate and go on from there to obtain a value for \( n \).

As an exercise in the use of Moment Generating Functions, we prove a simple special case of the Central Limit Theorem:

**Theorem.** If \( x \) is normally distributed with mean \( \mu \) and standard deviation \( \sigma \), and a random sample of size \( n \) is drawn, then the sample mean \( \bar{X} \) will be normally distributed with mean \( \mu \) and standard deviation \( \sigma/\sqrt{n} \). [Hoel, p. 139]
4.2. THE $\chi^2$ DISTRIBUTION.

Proof: since the sample values are random, independent, identically distributed and $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$, we can use the Moment Generating Function:

$$M_{\bar{x}}(\theta) = M_{(x_1 + \cdots + x_n)/n}(\theta) = M_{x_1 \cdots x_n}(\theta) = [M_x(\theta/n)]^n.$$  

It is trivial to check that, for a normally distributed random variable with mean $\mu$ and standard deviation $\sigma$, $M_x(\theta) = e^{\mu \theta + \frac{1}{2} \sigma^2 \theta^2}$. Replacing $\theta$ with $\theta/2$, we obtain:

$$M_{\bar{x}}(\theta) = [e^{\mu \theta + \frac{1}{2} \frac{\theta^2}{n} \sigma^2}]^n = e^{\mu \theta + \frac{1}{2} \frac{\sigma^2}{n}}^n,$$

which is the generating function of a normally distributed random variable with the desired mean and standard deviation.

We now turn to the actual distributions.

4.2 The $\chi^2$ distribution.

There are a number of situations where various components of an experimental outcome are independently distributed. The (square root of the) sum of their squares would then give a radial summary of their distribution. In case the components are normally distributed, we can obtain the explicit form of such a distribution.

Let the variables $x_1, x_2, \ldots, x_n$ denote a random sample from a normal population with mean 0 and variance 1. Let $w \equiv \sum_{i=1}^n x_i^2$. Then

$$M_w(\theta) = M_{(x_1^2 + \cdots + x_n^2)}(\theta) = M_{x_1^2}(\theta) \cdots M_{x_n^2}(\theta) = [M_2(\theta)]^n = \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\theta x^2} e^{-\frac{x^2}{2}} \, dx\right]^n.$$

The change of variable $y = x \sqrt{1 - 2\theta}$ allows us to transform the expression in the square brackets into $(1 - 2\theta)^{n/2}$, giving, finally $M_w(\theta) = (1 - 2\theta)^{n/2}$, a rather simple moment Generating Function. This will determine the probability distribution of $w$ uniquely, but we have no real clue as to what it is... As it turns out, the desired Probability Density Function is given by:

$$f(\chi^2) = \frac{(\chi^2)^{\nu/2-1} e^{-\chi^2/2}}{2^{\nu/2} \Gamma\left(\frac{\nu}{2}\right)}$$

where $\nu$, the number of independent random variables in $w$, is called the number of degrees of freedom of the distribution. One can now show, by some simple manipulations involving the Moment Generating Function associated with this distribution, that it has the same moment generating function we computed above, and is thus the desired distribution. We have:

Theorem: If $x$ is normally distributed with 0 mean and unit variance, the sum of squares of $n$ random sample values of $x$ has a $\chi^2$ distribution with $n$ degrees of freedom.
An Application. We want to calculate the probability that a rocket for a fireworks display designed to burst at a specified point in space will have a radial error of at most 100 m, if it is assumed that the $x$, $y$ and $z$ errors of a coordinate system with the origin at the specified point are independently normally distributed with a common standard deviation of 50 m. Let $x_1 = x/50$, $y_1 = y/50$ and $z_1 = z/50$. This guarantees that $x_1$, $y_1$ and $z_1$ are independent normally distributed random variables with zero means and unit variances. We have

$$P\{\sqrt{x^2 + y^2 + z^2} < 100\} = P\{x^2 + y^2 + z^2 < 100^2\} = P\{x_1^2 + y_1^2 + z_1^2 < 1\} = \frac{1}{2^{3/2} \Gamma(3/2)} \int_0^1 w^{1/2} e^{-w/2} \, dw$$

A numerical evaluation via Maple V gives

$$> \text{evalf}(1/(2^{-3/2} \text{GAMMA}(3/2))) \times \text{evalf}(\int \text{int}(w^{-1/2} \times \text{exp}(1)^{-w/2}), w=0..4));$$

$$0.7385358700$$

The $\chi^2$ distribution will be used extensively in testing goodness of fit: when experimental data will be used to determine the probability distribution as an elementary function. We finish with the statements of some further useful results about $\chi^2$ distributions.

**Theorem:** If $x$ is normally distributed with variance $\sigma^2$ and $s^2$ is the sample variance based on a random sample of size $n$, then $n s^2 / \sigma^2$ has a $\chi^2$ distribution with $n - 1$ degrees of freedom.

**Theorem:** If $\chi_1^2$ and $\chi_2^2$ possess independent $\chi^2$ distributions with $\nu_1$ and $\nu_2$ degrees of freedom, respectively, then $\chi_1^2 + \chi_2^2$ will possess a $\chi^2$ distribution with $\nu_1 + \nu_2$ degrees of freedom.

**Theorem:** If $n_1, \ldots, n_k$ and $e_1, \ldots, e_k$ are the observed and expected frequencies, respectively, for the $k$ possible outcomes of an experiment that is performed $n$ times, then, as $n$ becomes infinite, the distribution of the quantity

$$\sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$$

will approach that of a $\chi^2$ variable with $k - 1$ degrees of freedom.

### 4.3 Student's t distribution.

When the sample size is small, one cannot be confident that, for example, the sample standard deviation and the population standard deviation can be used interchangeably.

**Theorem:** If $u$ is normally distributed with zero mean and unit variance and $\nu^2$ has a $\chi^2$ distribution with $\nu$ degrees of freedom, and $u$ and $\nu$ are independently distributed, then the variable

$$t = \frac{u \sqrt{\nu}}{\nu}$$
4.4. THE F DISTRIBUTION.

has a Student's t distribution with $\nu$ degrees of freedom given by

$$f(t) = c \left( 1 + \frac{t^2}{\nu} \right)^{-(\nu+1)/2},$$

where

$$c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu} \Gamma\left(\frac{\nu}{2}\right)}.$$

**An Application: Confidence Limits for a Mean.** Let $x$ be normally distributed with mean $\mu$ and variance $\sigma^2$. Let $\bar{x}$ and $s^2$ be their sample estimates based on a random sample of size $n$. Then

$$u \equiv \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

and

$$v^2 \equiv \frac{n s^2}{\sigma^2}$$

satisfy the requirements of $u$ and $v$ in the theorem above. Recall that $s^2$ has $n - 1$ degrees of freedom, so that

$$t \equiv \frac{(\bar{x} - \mu) \sqrt{n - 1}}{s}$$

possesses a $t$ distribution with $n - 1$ degrees of freedom. If $t_{0.05}$ represents the value of $t$ such that the probability is 0.05 that $|t| > t_{0.05}$, the the probability is 0.95 that

$$\left| \frac{(\bar{x} - \mu) \sqrt{n - 1}}{s} \right| < t_{0.05}$$

or that

$$\bar{x} - t_{0.05} \frac{s}{\sqrt{n - 1}} < \mu < \bar{x} + t_{0.05} \frac{s}{\sqrt{n - 1}}$$

4.4 The F Distribution.

If one wishes to use the Student's t distribution to determine whether the means of two populations are the same, by taking samples, it turns out that the assumption that the population variances be the same is crucial to the applicability of the technique. It would thus be quite useful to have a frequency function that can be used for testing the equality of two variances.

**Theorem:** Let $u$ and $v$ possess independent $\chi^2$ distributions with $\nu_1$ and $\nu_2$ degrees of freedom, respectively, then

$$F = \frac{u}{v}$$

where
has the $F$ distribution with $\nu_1$ and $\nu_2$ degrees of freedom given by

$$f(F) = \frac{\nu_1^{\nu_1/2} \nu_2^{\nu_2/2} \Gamma \left( \frac{\nu_1 + \nu_2}{2} \right)}{\Gamma \left( \frac{\nu_1}{2} \right) \Gamma \left( \frac{\nu_2}{2} \right)} F^{(1/2)(\nu_1-2)} (\nu_2 + \nu_1 F)^{-(1/2)(\nu_1+\nu_2)}$$

[Hoel, p. 285]