Some Special Distributions
(Hogg Chapter Three)

STAT 405-01: Mathematical Statistics I [*]

Fall Semester 2013

Contents

[*] Copyright 2013, John T. Whelan, and all that

Tuesday 8 October 2013
– Read Section 3.1 of Hogg

1 Binomial & Related Distributions

1.1 The Binomial Distribution

Recall our first example of a discrete random variable, the number of heads on three flips of a fair coin. In that case, we were able to count the number of outcomes in each event, e.g., there were three ways to get two heads and a tail: $\{HHT, HTH, THH\}$, and each of the eight possible outcomes had the same probability, $\frac{1}{8} = 0.125$. If we consider a case where the coin is weighted so that each flip has a 60% chance of being a head and only a 40% chance of being a tail, then the probability of each of the ways to get two heads and a tail is:

$$P(HHT) = (0.60)(0.60)(0.40) = 0.144 \tag{1.1a}$$
$$P(HTH) = (0.60)(0.40)(0.60) = 0.144 \tag{1.1b}$$
$$P(THH) = (0.40)(0.60)(0.60) = 0.144 \tag{1.1c}$$

In each case, the probability is $(0.60)^2(0.40)^1 = 0.144$, so if X is the number of heads in three coin flips, $P(X = 2) = 0.144 + 0.144 + 0.144 = 0.432$. We can work out all of the probabilities in this way:

#H	outcomes	#	prob(each outcome)	tot prob
0	$\{TTT\}$	1	$(0.40)^3 = 0.064$	0.064
1	$\{TTH, THT, HTT\}$	3	$(0.60)(0.40)^2 = 0.096$	0.288
2	$\{THH, HTH, HHT\}$	3	$(0.60)^2(0.40) = 0.144$	0.432
3	$\{TTT\}$	1	$(0.60)^3 = 0.216$	0.216

This is an example of a *binomial random variable.* We have a set of n identical, independent "trials", experiments which could each turn out one of two ways. We call one of those possible results "success" (e.g., heads) and the other "failure" (e.g., tails). The probability of success on a given trial is some number $p \in [0, 1]$, so the non-negative integer n and real number p are parameters of the distribution. The random variable X is the number of successes in the n trials. Evidently, the possible values for X are $0, 1, 2, \ldots, n$. The probability of x successes is

$$p(x) = P(X = x) = (\#\text{ of outcomes})(\text{prob of each outcome}) \tag{1.2}$$

We can generalize the discussion above to see that the probability for any sequence containing x successes, and therefore $n - x$ failures is $p^x(1-p)^{n-x}$. In the simple example we just listed all of the outcomes in a particular event, but that won't be practical in general. Instead we fall back on a result from combinatorics. The number of ways of choosing x out of the n trials, which we call "n choose x", is

$$\binom{n}{x} = \frac{(n)(n-1)(n-2)\ldots(n-x+1)}{(x)(x-1)(x-2)\ldots(1)} = \frac{n!}{(n-x)!x!} \tag{1.3}$$

As a reminder, consider the number of distinct poker hands, consisting of five cards chosen from the 52-card deck. If we consider first the cards dealt out in order on the table in front of us, there are 52 possibilities for the first card, 51 for the second, 50 for the third, 49 for the fourth and 48=52-5+1 for the fifth. So the number of ways five cards could be dealt to us in order is $(52)(51)(50)(49)(48) = \frac{52!}{47!}$. But that is overcounting the number of distinct poker hands, since we can pick up the five cards and rearrange them anyway we like. Since we have 5 choices for the first card, 4 for the second, 3 for the third, 2 for the fourth and 1 for the fifth, any five-card poker hand can be rearranged $(5)(4)(3)(2)(1) = 5! = 120$ different ways. So in the list of $\frac{52!}{47!}$ ordered poker hands, each unordered hand is represented 5! = 120 times. Thus the number of unordered poker hands is

$$\binom{52}{5} = \frac{52!}{47!5!} = \frac{(52)(51)(50)(49)(48)}{(5)(4)(3)(2)(1)} \tag{1.4}$$

returning to the binomial distribution, this makes the pmf

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n \tag{1.5}$$

1.1.1 Some advice on calculating $\binom{n}{k}$

There is some art to calculating the binomial coëfficient

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{1.6}$$

For one thing, you basically never want to actually calculate the factorials. For example, consider $\binom{100}{2}$. It's easy to calculate if we write

$$\binom{100}{2} = \frac{(100)(99)}{(2)(1)} = (50)(99) = 4950 \tag{1.7}$$

On the other hand, if we write

$$\binom{100}{2} = \frac{100!}{98!2!} \tag{1.8}$$

we find that 100! and 98! are enormous numbers which will overflow calculators, single precision programs, etc.

A few identities which make these calculations easier:

$$\binom{n}{n-k} = \frac{n!}{k!(n-k)!} = \frac{n}{k} \tag{1.9a}$$

$$\binom{n}{0} = \binom{n}{n} = \frac{n!}{n!0!} = 1 \tag{1.9b}$$

$$\binom{n}{1} = \binom{n}{n-1} = \frac{n!}{(n-1)!1!} = n \tag{1.9c}$$

Note that this uses the definition that 0! = 1.

If k is small, you can write out the fraction, cancelling out $n-k$ factors to leave k factors in the numerator and k factors (including 1) in the denominator. (If $n-k$ is small, you cancel out k factors to leave $n-k$.) If k is really small you can use Pascal's triangle

$$\begin{array}{lccccccccc}
n=0\text{:} & & & & & 1 & & & & \\
n=1\text{:} & & & & 1 & & 1 & & & \\
n=2\text{:} & & & 1 & & 2 & & 1 & & \\
n=3\text{:} & & 1 & & 3 & & 3 & & 1 & \\
n=4\text{:} & 1 & & 4 & & 6 & & 4 & & 1
\end{array}$$

where element k of row n is $\binom{n}{k}$ and is created (for $n > 1$) by the recursion relation

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \tag{1.10}$$

I mention this mostly as a reminder that the binomial coëfficient is what appears in the binomial expansion

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \tag{1.11}$$

This can be used to show the binomial pmf is properly normalized:

$$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + [1-p])^n = 1^n = 1 \tag{1.12}$$

As a curiosity, note that sometimes a general formula [such as the recursion relation (1.10)] might tell us to calculate some sort of impossible binomial coëfficient, like $\binom{5}{-1}$ or $\binom{4}{6}$. Logically, it would seem reasonable to declare these to be zero, since there are no ways e.g., to pick six different items from a list of four. In fact, there's a sense in which the definition with factorials extends to this. For example,

$$\binom{3}{4} = \frac{3!}{4!(-1)!} \tag{1.13}$$

Now, the factorial of a negative number is not really defined, but if we try to extend the definition, it turns out to be, in some sense, infinite. We will, in the near future, define something known as the gamma function, usually defined by the integral

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt \tag{1.14}$$

It's easy to do the integral for $\alpha = 1$ and show that $\Gamma(1) = 1$ Using either integration by parts or parametric differentiation, we can show that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, from which you can show that for non-negative integer n, $\Gamma(n+1) = n!$. If we try to use the recursion relation to find $\Gamma(0)$, which would be $(-1)!$, we'd get

$$1 = \Gamma(1) = 0\Gamma(0) \tag{1.15}$$

which is not possible if $\Gamma(0)$ is finite. So it makes some sense to replace $\frac{1}{\Gamma(0)}$ with zero. Proceeding in this way you can show that $\Gamma(\alpha)$ also blows up if α is any negative integer. (We can also use the Gamma function to extend the definition of the factorial to non-integer values, but that's a matter for another time.)

Finally, there's the question of how to calculate binomial probabilities when x and $n-x$ are really large. At this point, you're going to be using a computer anyway. There are a few tricks for dealing with the large factorials, primarily by calculating $\ln p(x)$ rather than $p(x)$, but many statistical software packages do a lot of the work for you. For instance, in R, you can calculate the binomial pdf and/or cdf with commands like

```
n<-50
p<-0.25
x<-0:n
px<-dbinom(x,n,p)
Fx<-pbinom(x,n,p)
```

or with python

```
import numpy
from scipy.stats import binom
n=50
p=0.25
x=numpy.arange(n+1)
px=binom.pmf(x,n,p)
Fx=binom.cdf(x,n,p)
```

1.1.2 Properties of the Binomial Distribution

Returning to the binomial distribution, we can find its moment generating function and use it to find the mean and variance relatively quickly:

$$M(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx}p(x) = \binom{n}{x}(pe^t)^x(1-p)^{n-x} = (pe^t+[1-p])^n \tag{1.16}$$

It is often easier to work with the logarithm of the mgf, known as the *cumulant generating function*, which you studied in Hogg problem 2.4.8:

$$\psi(t) = \ln M(t) = n\ln(pe^t + [1-p]) \tag{1.17}$$

We use the chain rule to take its derivative:

$$\psi'(t) = \frac{npe^t}{pe^t + [1-p]} \tag{1.18}$$

from which we get the mean

$$E(X) = \psi'(0) = \frac{np}{p + [1-p]} = np \tag{1.19}$$

and differentiating again gives

$$\psi''(t) = \frac{npe^t}{pe^t + [1-p]} - \frac{np^2e^{2t}}{(pe^t + [1-p])^2} \tag{1.20}$$

from which we get the variance.

$$\mathrm{Var}(X) = \psi''(0) = np - np^2 = np(1-p) \tag{1.21}$$

These should be familiar results (the mean is sort of obvious if you think about it), but deriving them directly requires tricky manipulations of the sums. With the mgf it's child's play.

1.2 The Multinomial Distribution

The binomial distribution can be thought of as a special case of the multinomial distribution, which consists of a set of n identical and independent experiments, each of which has k possible outcomes, with probabilities $p_1, p_2, \ldots, p_k$, with $p_1 + p_2 + \cdots + p_k = 1$. (For the binomial case, $k = 1$, p_1 becomes p, and p_2 is $1-p$.) Define random variables X_1, X_2, $\ldots X_{k-1}$ which are the number of experiments out of the n total with each outcome. (We could also define X_k, but it is completely determined by the others as $X_k = n - X_1 - X_2 - \ldots - X_{k-1}$. The probability of any particular sequence of outcomes of which x_1 are of the first sort, x_2 of the same sort, etc, with $x_1 + x_2 + \cdots + x_k = n$ is

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{1.22}$$

How many different such outcomes are there? Well, there are $\binom{n}{x_1} = \frac{n!}{x_1!(n-x_1)!}$ ways to pick the x_1 experiments with the first outcome. Once we've done that, there are $n - x_1$ possibilities from which to choose the x_2 outcomes of the second sort, so there are $\binom{n-x_1}{x_2} = \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!}$ ways to do that, or a total of

$$\begin{aligned}\binom{n}{x_1}\binom{n-x_1}{x_2} &= \frac{n!}{x_1!(n-x_1)!}\frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} \\ &= \frac{n!}{x_1!x_2!(n-x_1-x_2)!} \end{aligned} \tag{1.23}$$

ways to pick the experiments that end up with the first two sorts of outcomes. Continuing in this way, we find the total number of outcomes of this sort to be

$$\begin{aligned}\binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\cdots\binom{n-x_1-x_2-\cdots-x_{k-1}}{x_k} \\ = \frac{n!}{x_1!x_2!\cdots x_k!}\end{aligned} \tag{1.24}$$

So the joint pdf for these multinomial random variables is

$$\begin{aligned}p(x_1, x_2, \ldots, x_{k-1}) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \\ x_1 = 0, 1, \ldots n; \quad x_2 = 0, 1, \ldots, n - x_1; \\ \cdots; \quad x_k = n - x_1 - x_2 - \cdots - x_{k-1}\end{aligned} \tag{1.25}$$

Thursday 10 October 2013 – Read Section 3.2 of Hogg

1.3 More on the Binomial Distribution

Recall that if we add two independent random variables X_1 and X_2 with mgfs $M_1(t) = E(e^{tX_1})$ and $M_2(t) = E(e^{tX_2})$, we can work out the distribution of their sum $Y = X_1 + X_2$ by finding its mgf:

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{t(X_1+X_2)}) = E(e^{tX_1}e^{tX_1}) \\ &= E(e^{tX_1})E(e^{tX_1}) = M_1(t)M_2(t) \end{aligned} \tag{1.26}$$

where we have used the fact that the expectation value of a product of functions of independent random variables is the product of their expectation values.

Now suppose we add independent binomial random variables: X_1 has n_1 trials with a probability of p (the notation introduced by Hogg calls this a $b(n_1, p)$) and X_2 has n_2 trials, also with a probability of p (which Hogg would call $b(n_2, p)$). Their mgfs are thus

$$M_1(t_1) = (pe^t + [1-p])^{n_1} \quad \text{and} \quad M_2(t_2) = (pe^t + [1-p])^{n_2} \tag{1.27}$$

If we call their sum $Y = X_1 + X_2$, its mgf is

$$M_Y(t) = M_1(t)M_2(t) = (pe^t + [1-p])^{n_1+n_2} \tag{1.28}$$

but this is just the mgf of a $b(n_1 + n_2, p)$ distribution. This important result also makes sense from the definition of the binomial distribution. Adding the number of successes in n_1 trials and in n_2 more trials is the same as counting the number of successes in $n_1 + n_2$ trials, as long as all the trials are independent, and the probability for success on each one is p.

1.4 The Negative Binomial Distribution

The binomial distribution is appropriate for describing a situation where the number of trials is set in advance. Sometimes, however, one decides to do as many trials as are needed to obtain a certain number of successes. The random variable in that situation would then be the number of trials that were required, or equivalently, the number of failures that occurred. This is know as a negative binomial random variable: given independent trials, each with probability p of success, X is the number of failures before the rth success. This probability is a little more involved to estimate. To get $X = x$, we have to have $r-1$ successes (and x failures) in the first $x+r-1$ trials, and then a success in the last trial, so the pmf is

$$p(x) = \binom{x+r-1}{r-1} p^{r-1}(1-p)^x p = \binom{x+r-1}{r-1} p^r (1-p)^x \tag{1.29}$$

We can have any (non-negative integer) number of failures, so the pmf is defined for $x = 0, 1, 2, \ldots$.

You will explore this distribution on the homework, and show that the mgf is

$$M(t) = p^r[1-(1-p)e^t]^{-r} \tag{1.30}$$

The factor in brackets is a binomial raised to a negative power, which is where the name "negative binomial distribution" comes from.

Note that in the special case $r = 1$, this is the number of failures before the first success, which is the *geometric distribution* which you've considered on a prior homework.

As an aside, the distinction between an experiment where you've decided in advance to do n trials, and one where you've decided to stop after r successes, leads to a complication in classical statistical inference called the *stopping problem*. The problem arises because in classical, or frequentist, statistical inference, you have to analyze the results of your experiment in terms of what would happen if you repeated the same experiment many times, and think about how likely results like you saw would be given various statistical models and parameter values. To answer such questions, you have to know what experiment you were planning to do, not just what observations you actually made. One of the advantages of Bayesian statistical inference, which asks the more direct question of how likely various models and parameter values are, given what you actually observed, is that you typically don't have to say what you'd do if you repeated the experiment.

1.5 The Hypergeometric Distribution

Both the binomial and negative binomial distribution result from repeated independent trials with the same probability p of success on each trial. This is a situation that's sometimes called *sampling with replacement*, since it's what would happen if you put, e.g., 35 red balls and 65 white balls in a bowl, drew one at random, noted its color, put it back, mixed up the balls, and repeated. Each time there would be a probability of $35/(35+65) = 0.35$ of drawing a red ball. If, on the other hand, you picked the ball out and didn't put it back, the probability of drawing a red ball on the next try would change. If the first ball was red, you'd have only a $34/99 \approx 0.3434$ chance of the next one being red (since there are now 34 red balls and still 65 white balls in the bowl), while if the first one was white, the probability for the second one to be red would go up to $35/99 \approx 0.3535$. This is called *sampling without replacement*. In practice, making some sort of huge tree diagram for all of the conditional probabilities, draw after draw, is impractical, so instead, if you want to know the probability to draw, say, three red balls out of ten, you count up the number of ways to do it. If you pick 10 balls out of a bowl containing 100, there are $\binom{100}{10} = \frac{100!}{90!10!}$ ways to do that. To count the number of those ways that have 3 red balls and 7 white balls, you need to count the number of ways to pick 3 of the 35 red balls, which is $\binom{35}{3} = \frac{35!}{32!3!}$ times the number of ways to pick 7 of the 65 white balls, which is $\binom{65}{7} = \frac{65!}{58!7!}$, so the probability of drawing exactly 3 red balls out of 10 when sampling without replacement from a bowl with 35 red balls out of 100 is

$$\frac{\binom{35}{3}\binom{65}{7}}{\binom{100}{10}} \tag{1.31}$$

In general, if there are N balls, of which D are red, and we draw n of them without replacement, the number of red balls in our sample will be a hypergeometric random variable X whose pmf is

$$p(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \tag{1.32}$$

As you might guess from the name, the mgf for a hypergeometric distribution is a hypergeometric function, which doesn't simplify things much. When D and $N - D$ are both large compared to n, the hypergeometric distribution reduces approximately to the binomial distribution. (Jaynes[1] uses the hypergeometric distribution in many of his examples, to avoid making the simplifying assumption of sampling with replacement.)

2 The Poisson Distribution

Consider the numerous statistical statements[2] you get like:

1. On average 118.3 people per day are killed in traffic accidents in the US
2. On average there are 367.2 gamma-ray bursts detected per year by orbiting satellites
3. During a rainstorm, an average of 929.4 raindrops falls on a square foot of ground each minute

[1] E. T. Jaynes, *Probability Theory: the Logic of Science* (Cambridge, 2003)

[2] I fabricated the last few significant figures in each of these numbers to produce a concrete example.

In each of these cases, the number of events, X, occurring in one representative interval is a discrete random variable with a probability mass function. The average number of events occurring is a parameter of this distribution, which Hogg writes as m. From the information given above, we only know the mean value of the distribution, $E(X) = m$. The extra piece of information that defines a Poisson random variable is that each of the discrete events to be counted is independent of the others.

2.1 Aside: Simulating a Poisson Process

To underline what this independence of events means, and to further to illustrate the kind of situation described by a Poisson distribution, consider the following thought experiment. Suppose there's a field divided into a million square-foot patches, and I'm told that there is an average of 4.5 beetles per square foot patch, with the location of each beetle independent of the others. If I wanted to simulate that situation, I could distribute 4,500,000 beetles randomly among the one million patches. I'd do this by placing the beetles one at a time, randomly choosing a patch (among the million available) for each beetle, without regard to where any of the other beetles were placed. The number of beetles in any one patch would then be pretty close to a Poisson random variable. (It wouldn't be exactly a Poisson rv, because the requirement of exactly 4,500,000 beetles total would break the independence of the different patches, but in the limit of infinitely many patches, and a corresponding total number of beetles, it would become an arbitrarily good approximation.) In fact, the numbers of beetles in each of the one million patches would make a good statistical ensemble for approximately describing the probability distribution for the Poisson random variable.

2.2 Poisson Process and Rate

These scenarios described by a Poisson random variable are often associated with some sort of a *rate*: 118.3 deaths per day, 367.2 GRBs per year, 929.4 raindrops per square foot per minute, 4.5 beetles per square foot. We talk about a Poisson process with an associated rate α (e.g., a number of events per unit time), and if we count all the events in a certain interval of size T (e.g., a specified duration of time), it is a Poisson random variable with parameter $m = \alpha T$. So for example, if the clicks on a Geiger counter are described by a Poisson process with a rate of 45 clicks per minute, the number of clicks in an arbitrarily chosen one-minute interval of time will be a Poisson random variable with $m = 45$. If the interval of time is 30 seconds (half a minute), the Poisson parameter will be $m = (45/\text{minute})(0.5\,\text{minute}) = 23.5$. If we count the clicks in four minutes, the number of clicks will be a Poisson random variable with $m = (45/\text{minute})(4\,\text{minutes}) = 180$. Note that m will always be a number, and will not contain a "per minute" or "per square foot", or anything. (That will be part of the rate α.)

2.3 Connection to Binomial Distribution and Derivation of pmf

The key information that lets us derive the pmf of a Poisson distribution is what happens if you break the interval up into smaller pieces. If there's a Poisson process at play, the number of events in each subinterval will also be a Poisson random variable, and they will all be independent of each other. So if we divide the day into 100 equal pieces of 14 minutes 24 seconds each, the number of traffic deaths in each is an independent Poisson random variable with a mean value of 11.83. Now, in practice this assumption will often not quite be true: the rate of traffic deaths is higher at some times of the day than others, some patches of ground may be more prone to be rained on because of wind patterns, etc, but we can imagine an idealized situation in which this subdivision works.

Okay, so how do we get the pmf? Take the interval in question (time interval, area of ground, or whatever) and divide it into n pieces. Each one of them will have an average of m/n events. If we make n really big, so that $m/n \ll 1$, the probability of getting one event in that little piece will be small, and the probability that two or more of them happen to occur in the same piece is even smaller, and we can ignore it to a good approximation. (We can always make the approximation better by making n bigger.) That means that the number of events in that little piece, call it Y, has a pmf of

$$p(Y = 0) \approx 1 - p \tag{2.1a}$$
$$p(Y = 1) = p \tag{2.1b}$$
$$p(Y > 1) \approx 0 \tag{2.1c}$$

In order to have $E(Y) = m/n$, the probability of an event occurring in that little piece has to be $p = m/n$.

But we have now described the conditions for a binomial distribution! Each of the n tiny sub-pieces of the interval is like a trial, a piece with an event is a success, and a piece with no event is a failure. So the pmf for this Poisson random variable must be the limit of a binomial distribution as the number of trials gets large:

$$\begin{aligned} p(x) &= \lim_{n\to\infty} \frac{n!}{x!(n-x)!}\left(\frac{m}{n}\right)^x \left(1-\frac{m}{n}\right)^{n-x} \\ &= \frac{m^x}{x!}\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n \frac{n!}{(n-x)!}\left(\frac{1/n}{1-m/n}\right)^x \end{aligned} \tag{2.2}$$

Now, the ratio of factorials is a product of x things:

$$\frac{n!}{(n-x)!} = n(n-1)(n-2)\cdots(n-x+1) \tag{2.3}$$

The last factor is of course also the product of x things, i.e., x identical copies of

$$\frac{1/n}{1-m/n} = \frac{1}{n-m} \ . \tag{2.4}$$

But that means the two of them together give you

$$\frac{n!}{(n-x)!}\left(\frac{1/n}{1-m/n}\right)^x = \frac{n}{n-m}\,\frac{n-1}{n-m}\,\frac{n-2}{n-m}\cdots\frac{n-x+1}{n-m} \tag{2.5}$$

which is the product of x fractions, each of which goes to 1 as n goes to infinity, so we can lose that factor in the limit and get

$$p(x) = \frac{m^x}{x!}\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = \frac{m^x}{x!}e^{-m} \ . \tag{2.6}$$

where we've used the exponential function

$$e^\alpha = \lim_{n\to\infty}\left(1+\frac{\alpha}{n}\right)^n \ ; \tag{2.7}$$

If we recall that the exponential function also has the Maclaurin series

$$e^\alpha = \sum_{k=0}^{\infty}\frac{\alpha^k}{k!} \tag{2.8}$$

we can see that the pmf

$$p(x) = \frac{m^x}{x!}e^{-m} \tag{2.9}$$

is normalized:

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty}\frac{m^x}{x!}e^{-m} = e^m e^{-m} = 1 \tag{2.10}$$

We can also use this series to get the mgf

$$M(t) = \sum_{x=0}^{\infty} e^{tx} \frac{m^x}{x!} e^{-m} = \frac{(me^t)^x}{x!} e^{-m} = e^{me^t} e^{-m} = e^{m(e^t-1)} \tag{2.11}$$

To find the mean and variance, it's useful once again to use the cumulant generating function $\psi(t) = \ln M(t)$:

$$\psi(t) = m(e^t - 1) \tag{2.12}$$

Differentiating gives us

$$\psi'(t) = me^t \tag{2.13}$$

and

$$\psi''(t) = me^t \tag{2.14}$$

so the mean is

$$E(X) = \psi'(0) = m \tag{2.15}$$

(which was the definition we started with) and the variance is

$$\mathrm{Var}(X) = \psi''(0) = m \tag{2.16}$$

Note that these are also the limits of the mean and variance of the binomial distribution.

We can also the mgf to show what happens when we add two Poisson random variables with means m_1 and m_2; the mgf of their sum is

$$M(t) = e^{m_1(e^t-1)} e^{m_2(e^t-1)} = e^{(m_1+m_2)(e^t-1)} \tag{2.17}$$

which is the mgf of a Poisson random variable with mean $m_1 + m_2$, so the sum of two Poisson rvs is itself a Poisson rv.

Tuesday 15 October 2013 – No Class (Monday Schedule)

Thursday 17 October 2013 – Read Section 3.3 of Hogg

3 Gamma and Related Distributions

3.1 Gamma (Γ) Distribution

We now consider our first family of continuous distributions, the Gamma distribution. A Gamma random variable has a pdf with two parameters $\alpha > 0$ and $\beta > 0$:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \qquad 0 < x < \infty \tag{3.1}$$

We sometimes say that such a random variable has a $\Gamma(\alpha, \beta)$ distribution. As we'll see, the β parameter just changes the scale of the distribution function, but α actually changes the shape. We will show how special cases of the Gamma distribution describe situations of interest, notably the exponential distribution with rate λ, which is $\Gamma(1, \frac{1}{\lambda})$, and the chi-squared distribution with r degrees of freedom (also known as $\chi^2(r)$), which is $\Gamma(\frac{r}{2}, 2)$.

For now, though, let's look at the general Gamma distribution. First of all, consider the constant $\frac{1}{\Gamma(\alpha)\beta^\alpha}$. The $\Gamma(\alpha)$ in the denominator is the Gamma function[3]

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du \tag{3.2}$$

which we introduced last week as a generalization of the factorial function, so that if n is a non-negative integer, $\Gamma(n+1) = n!$.

[3] Note that in an unfortunate collision of notation, $\Gamma(\alpha)$ is a function which returns a number for each value of α, while $\Gamma(\alpha, \beta)$ is a label we use to refer to a distribution corresponding to a choice of α and β.

That constant is just what we need to make sure the distribution function is normalized:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/\beta}\,dx \\ &= \frac{1}{\Gamma(\alpha)}\int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta}\,\frac{dx}{\beta} \\ &= \frac{1}{\Gamma(\alpha)}\int_0^\infty u^{\alpha-1}e^{-u}\,du = \frac{1}{\Gamma(\alpha)}\Gamma(\alpha) = 1\end{aligned} \tag{3.3}$$

where we have made the change of variables $u = x/\beta$. The cdf for the Gamma distribution is

$$\begin{aligned}F(x) &= \int_{-\infty}^{x} f(y)\,dy = \frac{1}{\Gamma(\alpha)}\int_0^x \left(\frac{y}{\beta}\right)^{\alpha-1} e^{-y/\beta}\,\frac{dy}{\beta} \\ &= \frac{1}{\Gamma(\alpha)}\int_0^{x/\beta} u^{\alpha-1}e^{-u}\,du\end{aligned} \tag{3.4}$$

which is usually written in terms of either the *incomplete gamma function*

$$\gamma(y;\alpha) = \int_0^y u^{\alpha-1}e^{-u}\,du \tag{3.5}$$

defined so that $\lim_{y\to\infty}\gamma(y;\alpha) = \Gamma(\alpha)$ or the *standard Gamma cdf*

$$F(y;\alpha) = \frac{1}{\Gamma(\alpha)}\int_0^y u^{\alpha-1}e^{-u}\,du \tag{3.6}$$

defined so that $\lim_{y\to\infty} F(y;\alpha) = 1$. One or both of these functions may be available in your favorite software package, or tabulated in a book. (For instance, $F(y;\alpha)$ is tabulated in the back of Devore, *Probability and Statistics for Engineering and the Sciences*, the book you presumably used for intro Probability and Statistics.) Hogg does not tabulate these, but it does include some values of the chi-squared cdf, which is a special case of the Gamma cdf.

Note that if we define a random variable $Y = X/\beta$, where X is a $\Gamma(\alpha,\beta)$ random variable, its pdf is

$$f_Y(y) = \frac{dP}{dy} = \frac{dP/dx}{dy/dx} = \frac{f_X(\beta y)}{1/\beta} = \frac{1}{\Gamma(\alpha)}y^{\alpha-1}e^{-y} \tag{3.7}$$

which is a $\Gamma(\alpha,1)$ distribution. This is what we mean when we say β is a scale parameter.

Now we turn to the mgf of the Gamma distribution, which is

$$\begin{aligned}M(t) = E(e^{tX}) &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/\beta}e^{tx}\,dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}\exp\left(-\left[\frac{1}{\beta}-t\right]x\right)\,dx\end{aligned} \tag{3.8}$$

If we require $t < \frac{1}{\beta}$ and make the substitution $u = (\frac{1}{\beta}-t)x$ so $x = \frac{\beta}{1-\beta t}\,u$, this becomes

$$M(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\left(\frac{\beta}{1-\beta t}\right)^\alpha \int_0^\infty u^{\alpha-1}e^{-u}\,du = (1-\beta t)^{-\alpha} \tag{3.9}$$

If we again construct the cumulant generating function

$$\psi(t) = \ln M(t) = -\alpha\ln(1-\beta t) \tag{3.10}$$

we find the derivative

$$\psi'(t) = \alpha\beta(1-\beta t)^{-1} \tag{3.11}$$

so the mean is

$$E(X) = \psi'(0) = \alpha\beta \tag{3.12}$$

and the second derivative

$$\psi''(t) = \alpha\beta^2(1-\beta t)^{-2} \tag{3.13}$$

so the variance is

$$\mathrm{Var}(X) = \psi''(0) = \alpha\beta^2 \tag{3.14}$$

which should be familiar results from elementary probability and statistics.

Even more useful than the mean and variance, the mgf can be used to show what happens when we add two independent Gamma random variables with the same scale parameter, say X_1 which is $\Gamma(\alpha_1, \beta)$ and X_2 which is $\Gamma(\alpha_2, \beta)$. The mgf of their sum is

$$M(t) = M_1(t)M_2(t) = (1-\beta t)^{-\alpha_1}(1-\beta t)^{-\alpha_2} = (1-\beta t)^{-(\alpha_1+\alpha_2)} \tag{3.15}$$

which we see is the mgf of a $\Gamma(\alpha_1+\alpha_2, \beta)$ random variable. So if you add Gamma rvs with the same scale parameter, their sum is another Gamma rv whose scale parameter is the sum of the individual scale parameters.

3.2 Exponential Distribution

Suppose we have a Poisson process with rate λ, so that the probability mass function of the number of events occurring in any period of time of duration t is

$$P(n \text{ events in } t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{3.16}$$

We can show that, if you start counting at some moment, the time you have to wait until k more events have occurred is a $\Gamma(k, \frac{1}{\lambda})$ random variable. In Hogg they show this by summing the Poisson distribution from 0 to $k-1$, but once we know what happens when you sum Gamma random variables, all we actually have to do is show that the waiting time for the *first* event is a $\Gamma(1, \frac{1}{\lambda})$ random variable. If you start waiting for the second event once the first has happened, that waiting time is another (independent) $\Gamma(1, \frac{1}{\lambda})$ random variable, and so forth. The waiting time for the kth event is thus the sum of k independent $\Gamma(1, \frac{1}{\lambda})$ random variables, which by the addition property we've just seen is a $\Gamma(k, \frac{1}{\lambda})$ random variable.

So we turn to the question of the waiting time for the first event, and return to the Poisson distribution. In particular, evaluating the pmf at $n = 0$ we get

$$P(\text{no events in } t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \tag{3.17}$$

Now pick some arbitrary moment and let X be the random variable describing how long we have to wait for the next event. If $X \leq t$ that means there are one or more events occurring in the interval of length t starting at that moment, so the probability of this is

$$P(X \leq t) = 1 - P(\text{no events in } t) = 1 - e^{-\lambda t} \tag{3.18}$$

But that is the definition of the cumulative distribution function, so

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x} \tag{3.19}$$

Note that it doesn't make sense for X to take on negative values, which is good, since $F(0) = 0$. This means that technically,

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases} \tag{3.20}$$

We can differentiate (with respect to x) to get the pdf

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases} \tag{3.21}$$

This is known as the *exponential* distribution, and if we recall that $\Gamma(0) = 1! = 1$, we see that this is indeed the Gamma distribution where $\alpha = 1$ and $\beta = \frac{1}{\lambda}$.

3.3 Chi-Square (χ^2) Distribution

We turn now to another special case of the Gamma distribution. If we set α to $\frac{r}{2}$, where r is some positive integer, and β to 2, we get the pdf

$$f(x) = \frac{1}{\Gamma(\frac{r}{2})2^{r/2}} x^{\frac{r}{2}-1} e^{-x/2} \qquad 0 < x < \infty \tag{3.22}$$

which is known as a chi-square distribution with r degrees of freedom, or $\chi^2(r)$. Note that if we add independent chi-square random variables, e.g., X_1 which is $\chi^2(r_1) \equiv \Gamma(\frac{r_1}{2}, 2)$ and X_2 which is $\chi^2(r_2) \equiv \Gamma(\frac{r_2}{2}, 2)$, their sum $X_1 + X_2$ is $\Gamma(\frac{r_1}{2} + \frac{r_2}{2}, 2) \equiv \chi^2(r_1 + r_2)$. So in particular the sum of r independent $\chi^2(1)$ random variables is a $\chi^2(r)$ random variable. Next week, we'll see that a $\chi^2(1)$ random variable is the square of a standard normal random variable, so a $\chi^2(r)$ is the sum of the squares of r independent standard normal random variables.

Note that Table II in Appendix C of Hogg has some percentiles (values of x for which $P(X < x)$ is some given value) for chi-square distributions with various numbers of degrees of freedom. This means that if you have a Gamma random variable which you can scale to be a chi-square random variable (which is basically possible if α is any integer or half-integer), you can get percentiles of that random variable.

3.4 Beta (β) Distribution

Please read about the β distribution in section 3.3 of Hogg.

Tuesday 22 October 2013 – Read Section 3.4 of Hogg

4 The Normal Distribution

A random variable X follows a normal distribution, also known as a Gaussian distribution, with parameters μ and $\sigma > 0$, known as $N(\mu, \sigma^2)$, if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.1}$$

To show that this is normalized, we take the integral

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\,dz \tag{4.2}$$

where we have made the substitution $z = (x-\mu)/\sigma$. The integral

$$I = \int_{-\infty}^{\infty} e^{-z^2/2}\,dz \tag{4.3}$$

is a bit tricky; there's no ordinary function whose derivative is $e^{-z^2/2}$, so we can't just do an indefinite integral and evaluate at the endpoints. But we can do the definite integral by writing

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy \end{aligned} \tag{4.4}$$

If we interpret this as a double integral in Cartesian coördinates, we can change to polar coördinates r and ϕ, and write

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}\,r\,dr\,d\phi = 2\pi \int_0^{\infty} e^{-r^2/2}\,r\,dr \\ &= -2\pi\, e^{-r^2/2}\Big|_0^{\infty} = 2\pi \end{aligned} \tag{4.5}$$

so

$$I = \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \sqrt{2\pi} \tag{4.6}$$

and the pdf does integrate to one.

To get the mgf, we have to take the integral

$$M(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + tx\right)\, dx \tag{4.7}$$

If we complete the square in the exponent, we get

$$\begin{aligned} -\frac{(x-\mu)^2}{2\sigma^2} + tx &= -\frac{1}{2\sigma^2}\left(x - [\mu + t\sigma^2]\right)^2 + \frac{1}{2\sigma^2}\left(2\mu t\sigma^2 + t^2\sigma^4\right) \\ &= -\frac{1}{2\sigma^2}\left(x - [\mu + t\sigma^2]\right)^2 + \mu t + \frac{t^2\sigma^2}{2} \end{aligned} \tag{4.8}$$

so

$$\begin{aligned} M(t) &= \frac{1}{\sigma\sqrt{2\pi}} e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x - [\mu + t\sigma^2])^2}{2\sigma^2}\right)\, dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = e^{\mu t + \frac{t^2\sigma^2}{2}} \end{aligned} \tag{4.9}$$

This means that the cumulant generating function is

$$\psi(t) = \ln M(t) = \mu t + \frac{\sigma^2 t^2}{2} \tag{4.10}$$

taking derivatives gives

$$\psi'(t) = \mu + \sigma^2 t \tag{4.11}$$

so the mean is

$$E(X) = \psi'(0) = \mu \tag{4.12}$$

and

$$\psi''(t) = \sigma^2 \tag{4.13}$$

so the variance is

$$\mathrm{Var}(X) = \psi''(0) = \sigma^2 \tag{4.14}$$

which means that the parameters μ and σ are the mean and standard deviation of the distribution, as their names suggest. Note that this is in some sense the "simplest" possible distribution with a given mean and variance. For a general random variable, since $\psi(0) = 0$, $\psi'(0) = E(X)$ and $\psi''(0) = \mathrm{Var}(X)$, the first few terms of the Maclaurin series for $\psi(t)$ must be

$$\psi(t) = t\, E(X) + \frac{t^2}{2}\, \mathrm{Var}(X) + \mathcal{O}(t^3) \tag{4.15}$$

Given a random variable X which follows a $N(\mu, \sigma^2)$ distribution, we can define $Z = \frac{X-\mu}{\sigma}$. Its pdf will be

$$f_Z(z) = \sigma f_X(\mu + z\sigma) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{4.16}$$

which is a $N(1, 0)$ distribution, also known as a *standard normal distribution.*

The cdf of a $N(\mu, \sigma)$ random variable will be

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(u-\mu)^2/(2\sigma^2)}\, du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2}\, dt \tag{4.17}$$

again, $e^{-t^2/2}$ is not the derivative of any known function, but it's useful enough that we define a function

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt \tag{4.18}$$

which is tabulated in lots of places. In terms of this, the cdf for a $N(\mu, \sigma^2)$ rv is

$$P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{4.19}$$

If we add two independent Gaussian random variables, X_1 and X_2, following $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions, respectively, their sum has the mgf

$$\begin{aligned} M(t) = M_1(t)M_2(t) &= \exp\left(t\mu_1 + \frac{t^2\sigma_1^2}{2}\right)\exp\left(t\mu_2 + \frac{t^2\sigma_2^2}{2}\right) \\ &= \exp\left(t(\mu_1+\mu_2) + \frac{t^2(\sigma_1^2+\sigma_2^2)}{2}\right) \end{aligned} \tag{4.20}$$

which is the mgf of a $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ distribution. (In general, if we add independent random variables, their sum has a mean which is the sum of the means and a variance which is the sum of the variances, but what's notable here is the sum obeys a normal distribution, so it's characterized only by its mean and variance.

Finally, suppose we have r independent Gaussian random variables $\{X_i\}$ with means μ_i and variances σ_i^2. Consider the combination

$$Y = \sum_{i=1}^{r}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \tag{4.21}$$

We can show that this obeys a $\chi^2(r)$ distribution

$$f_Y(y) = \frac{1}{\Gamma(r/2)2^{r/2}} y^{\frac{r}{2}-1} e^{-y/2} \tag{4.22}$$

As we saw last time, the sum of r independent $\chi^2(1)$ random variables is a $\chi^2(r)$ random variable, so all we need to do is show that if X is $N(\mu, \sigma^2)$, so that $Z = \frac{X-\mu}{\sigma}$ is $N(0,1)$, $Y = \frac{(X-\mu)^2}{\sigma} = Z^2$ is $\chi^2(1)$. Now, since

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty \tag{4.23}$$

we can't quite use the usual formalism for transformation of pdfs, since the transformation $Y = Z^2$ is not invertible. But since $f_Z(-z) = f_Z(z)$ it's not hard to see that if we define a rv $W = |Z|$, it must have a pdf

$$f_W(w) = f_Z(-w) + f_Z(w) = 2f_Z(z) = \frac{2}{\sqrt{2\pi}} e^{-w^2/2} \quad 0 < w < \infty \tag{4.24}$$

and then we can use the transformation $y = w^2$, $w = y^{1/2}$ to work out

$$\begin{aligned} f_Y(y) &= \frac{dP}{dy} = \frac{dP}{dw}\frac{dw}{dy} = \frac{1}{2} y^{-1/2} f_W(y^{1/2}) \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \qquad 0 < y < \infty \end{aligned} \tag{4.25}$$

If we recall the $\chi^2(1)$ pdf from last time, it was

$$f(y) = \frac{1}{\Gamma(1/2)2^{1/2}} y^{-1/2} e^{-y/2} \quad 0 < y < \infty \tag{4.26}$$

so the pdf of $Y = Z^2$ is the $\chi^2(1)$ pdf, if the value of $\Gamma(1/2)$ is $\sqrt{\pi}$. Now, if we think about it, that has to be the case, in order for the two pdfs to be normalized, but we can work out the value directly. Recall the Gamma function

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt \tag{4.27}$$

which is a finite positive number for any $\alpha > 0$. (For positive integer n, we know that $\Gamma(n) = (n-1)!$.) Thus

$$\Gamma(1/2) = \int_0^\infty t^{-1/2} e^{-t}\, dt \tag{4.28}$$

We can show that the integral is well-behaved at the lower limit of $t = 0$, and evaluate it, by changing variables to $u = \sqrt{2t}$ so that $t = u^2/2$ and $du = 2^{1/2} t^{-1/2}\, dt$; thus

$$\Gamma(1/2) = \frac{1}{\sqrt{2}} \int_0^\infty e^{-u^2}\, du = \frac{1}{\sqrt{2}} \frac{\sqrt{2\pi}}{2} = \sqrt{\pi} \tag{4.29}$$

as expected. We've used the symmetry of the integrand to say that

$$\int_0^\infty e^{-u^2}\,du = \frac{1}{2}\int_{-\infty}^\infty e^{-u^2}\,du = \frac{\sqrt{2\pi}}{2} \tag{4.30}$$

Thursday 24 October 2013 – Read Section 3.5 of Hogg

5 Multivariate Normal Distribution

5.1 Linear Algebra: Reminders and Notation

If $\mathbf{A}$ is an $m \times n$ matrix:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \tag{5.1}$$

and $\mathbf{B}$ is an $n \times p$ matrix,

$$\mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix} \tag{5.2}$$

then their product $\mathbf{C} = \mathbf{AB}$ is an $m \times p$ matrix as shown in Figure 1 so that $C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$.

If $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$ is an $n \times m$ matrix with elements $B_{ij} = A_{ji}$:

$$\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nm} \end{pmatrix} = \mathbf{B} = \mathbf{A}^{\mathrm{T}} = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{mn} \end{pmatrix} \tag{5.4}$$

If $\mathbf{v}$ is an n-element column vector (which is an $n \times 1$ matrix) and $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{w} = \mathbf{Av}$ is an m-element column vector (i.e., an $m \times 1$ matrix):

$$\begin{aligned} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \mathbf{w} = \mathbf{Av} &= \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \\ &= \begin{pmatrix} A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n \\ A_{21}v_1 + A_{22}v_2 + \cdots + A_{2n}v_n \\ \vdots \\ A_{m1}v_1 + A_{m2}v_2 + \cdots + A_{mn}v_n \end{pmatrix} \end{aligned} \tag{5.5}$$

so that $w_i = \sum_{j=1}^{n} A_{ij}v_j$.

If $\mathbf{u}$ is an n-element column vector, then $\mathbf{u}^{\mathrm{T}}$ is an n-element row vector (a $1 \times n$ matrix):

$$\mathbf{u}^{\mathrm{T}} = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} \tag{5.6}$$

If $\mathbf{u}$ and $\mathbf{v}$ are n-element column vectors, $\mathbf{u}^{\mathrm{T}}\mathbf{v}$ is a number, known as the *inner product*:

$$\begin{aligned} \mathbf{u}^{\mathrm{T}}\mathbf{v} &= \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \\ &= u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{i=1}^{n} u_iv_i \end{aligned} \tag{5.7}$$

If $\mathbf{v}$ is an m-element column vector, and $\mathbf{w}$ is an n-element

$$\mathbf{C} = \mathbf{AB} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mp} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix}$$

$$= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} + \cdots + A_{1n}B_{n1} & A_{11}B_{12} + A_{12}B_{22} + \cdots + A_{1n}B_{n2} & \cdots & A_{11}B_{1p} + A_{12}B_{2p} + \cdots + A_{1n}B_{np} \\ A_{21}B_{11} + A_{22}B_{21} + \cdots + A_{2n}B_{n1} & A_{21}B_{12} + A_{22}B_{22} + \cdots + A_{2n}B_{n2} & \cdots & A_{21}B_{1p} + A_{22}B_{2p} + \cdots + A_{2n}B_{np} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}B_{11} + A_{m2}B_{21} + \cdots + A_{mn}B_{n1} & A_{m1}B_{12} + A_{m2}B_{22} + \cdots + A_{mn}B_{n2} & \cdots & A_{m1}B_{1p} + A_{m2}B_{2p} + \cdots + A_{mn}B_{np} \end{pmatrix} \tag{5.3}$$

Figure 1: Expansion of the product $\mathbf{C} = \mathbf{AB}$ to show $C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$.

column vector, $\mathbf{A} = \mathbf{v}\mathbf{w}^{\mathrm{T}}$ is an $m \times n$ matrix

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} = \mathbf{A} = \mathbf{v}\mathbf{w}^{\mathrm{T}}$$
$$= \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \cdots & w_m \end{pmatrix} = \begin{pmatrix} v_1w_1 & v_1w_2 & \cdots & v_1w_n \\ v_2w_1 & v_2w_2 & \cdots & v_2w_n \\ \vdots & \vdots & \ddots & \vdots \\ v_mw_1 & v_mw_2 & \cdots & v_mw_n \end{pmatrix} \tag{5.8}$$

so that $A_{ij} = v_i w_j$.

If $\mathbf{M}$ and $\mathbf{N}$ are $n \times n$ matrices, the determinant $\det(\mathbf{MN}) = \det(\mathbf{M})\det(\mathbf{N})$.

If $\mathbf{M}$ is an $n \times n$ matrix (known as a square matrix), the inverse matrix $\mathbf{M}^{-1}$ is defined by $\mathbf{M}^{-1}\mathbf{M} = \mathbf{1}_{n\times n} = \mathbf{M}\mathbf{M}^{-1}$ where $\mathbf{1}_{n\times n}$ is the identity matrix

$$\mathbf{1}_{n\times n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{5.9}$$

If $\mathbf{M}^{-1}$ exists, we say $\mathbf{M}$ is invertible.

If $\mathbf{M}$ is a real, symmetric $n \times n$ matrix, so that $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$, i.e., $M_{ji} = M_{ij}$, there is a set of n orthonormal *eigenvectors* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ with real eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, so that $\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Orthonormal means

$$\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{5.10}$$

where we have introduced the Kronecker delta symbol δ_{ij}. The eigenvalue decomposition means

$$\mathbf{M} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{5.11}$$

The determinant is $\det(\mathbf{M}) = \prod_{i=1}^{n} \lambda_i$. If none of the eigenvalues $\{\lambda_i\}$ are zero, $\mathbf{M}$ is invertible, and the inverse matrix is

$$\mathbf{M}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{5.12}$$

If all of the eigenvalues $\{\lambda_i\}$ are positive, we say $\mathbf{M}$ is positive definite. If none of the eigenvalues $\{\lambda_i\}$ are negative, we say $\mathbf{M}$ is positive semi-definite.

5.2 Special Case: Independent Gaussian Random Variables

Before considering the general multivariate normal distribution, consider the case of n independent normally-distributed random variables $\{X_i\}$ with means $\{\mu_i\}$ and variances $\{\sigma_i\}$. The pdf for X_i, the ith random variable, is

$$f_i(x_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{5.13}$$

and its mgf is

$$M_i(t_i) = \exp\left(t_i\mu_i + \frac{1}{2}t_i^2\sigma_i^2\right) \tag{5.14}$$

If we consider the random variables X_i to be the elements of a random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{5.15}$$

its expectation value is

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \tag{5.16}$$

and its variance-covariance matrix is

$$\operatorname{Cov}(\mathbf{X}) = \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \tag{5.17}$$

which is diagonal because the different X_is are independent of each other and therefore have zero covariance. We can thus write the joint mgf for these random variables as

$$\begin{aligned} M(\mathbf{t}) &= \prod_{i=1}^{n} M_i(t_i) = \exp\left(\sum_{i=1}^{n}\left[t_i\mu_i + \frac{1}{2}t_i^2\sigma_i^2\right]\right) \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{t}\right) \end{aligned} \tag{5.18}$$

We can also write the joint pdf

$$f(\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i) = \frac{1}{\sqrt{\prod_{i=1}^{n} 2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right) \tag{5.19}$$

in matrix form if we consider a few operations on the matrix $\mathbf{\Sigma}$. First, since it's a diagonal matrix, its determinant is just the product of its diagonal entries:

$$\det \mathbf{\Sigma} = \prod_{i=1}^{n} \sigma_i^2 \tag{5.20}$$

and, for that matter,

$$\det(2\pi\boldsymbol{\Sigma}) = \prod_{i=1}^{n} 2\pi\sigma_i^2 \tag{5.21}$$

Also, we can invert the matrix to get

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{pmatrix} \tag{5.22}$$

so

$$\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{5.23}$$

which makes the pdf for the random vector X

$$f(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \tag{5.24}$$

The generalization from n independent normal random variables to an n-dimensional multivariate normal distribution is to use the same matrix form for $M(\mathbf{t})$ and just to replace $\boldsymbol{\Sigma}$, which was a diagonal matrix with positive diagonal entries, with a general symmetric positive semi-definite matrix. So one change is to allow $\boldsymbol{\Sigma}$ to have off-diagonal entries, and another is to allow it to have zero eigenvalues. If $\boldsymbol{\Sigma}$ is positive definite, i.e., its eigenvalues are all positive, we can use the matrix expression for the pdf $f(\mathbf{x})$ as well. As we'll see, if $\boldsymbol{\Sigma}$ has some zero eigenvalues, we won't be able to define a pdf for the random vector $\mathbf{X}$.

5.3 Multivariate Distributions

Recall that a set of n random variables X_1, X_2, $\ldots X_n$ can be combined into a *random vector*

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{5.25}$$

with expectation value

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \tag{5.26}$$

and variance-covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} &= \operatorname{Cov}(\mathbf{X}) = E([\mathbf{X} - \boldsymbol{\mu}]^{\mathrm{T}}[\mathbf{X} - \boldsymbol{\mu}]) \\ &= \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_1, X_2) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_1, X_n) & \operatorname{Cov}(X_2, X_n) & \cdots & \operatorname{Var}(X_n) \end{pmatrix} \end{aligned} \tag{5.27}$$

The variance-covariance matrix must be positive semi-definite, i.e., have no negative eigenvalues. To see why that is the case, let $\{\lambda_i\}$ be the eigenvalues and $\{\mathbf{v}_i\}$ be the orthonormal eigenvectors, so that $\boldsymbol{\Sigma} = \sum_{i=1}^{n} {\mathbf{v}^{\mathrm{T}}}_i \lambda_i \mathbf{v}_i$. For each i define a random variable $\mathcal{X}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{X}$. It has mean $E(\mathcal{X}_i) = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}$ and variance

$$\begin{aligned} \operatorname{Var}(\mathcal{X}_i) &= E([\mathbf{v}_i^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu})]^2) = E[\mathbf{v}_i^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{v}_i] \\ &= \mathbf{v}_i^{\mathrm{T}} E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}]\mathbf{v}_i = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}_i = \lambda_i \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = \lambda_i \end{aligned} \tag{5.28}$$

Since the variance of a random variable must be non-negative, $\mathbf{\Sigma}$ cannot have any negative eigenvalues.

Incidentally, we can see that the different random variables $\{\mathcal{X}_i\}$ are uncorrelated, since

$$\begin{aligned}\operatorname{Cov}(\mathcal{X}_i,\mathcal{X}_j) &= E[\mathbf{v}_i^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{v}_j] \\ &= \mathbf{v}_i^{\mathrm{T}}E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}]\mathbf{v}_j = \mathbf{v}_i^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = \lambda_i\delta_{ij}\end{aligned} \tag{5.29}$$

Remember that $\operatorname{Cov}(\mathcal{X}_i,\mathcal{X}_j)=0$ does not necessarily imply that $\mathcal{X}_i$ and $\mathcal{X}_j$ are independent. (It will, however, turn out to be the case for normally distributed random variables.)

Note that if we assemble the $\{\mathcal{X}_i\}$ into a column vector

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix}\mathcal{X}_1\\ \mathcal{X}_2\\ \vdots\\ \mathcal{X}_n\end{pmatrix} = \begin{pmatrix}\mathbf{v}_1^{\mathrm{T}}\\ \mathbf{v}_2^{\mathrm{T}}\\ \vdots\\ \mathbf{v}_n^{\mathrm{T}}\end{pmatrix}\mathbf{X} = \mathbf{\Gamma}\mathbf{X} \tag{5.30}$$

where the matrix $\mathbf{\Gamma}$ is made up out of the components of the orthonormal eigenvectors $\{\mathbf{v}_i\}$:

$$\mathbf{\Gamma} = \begin{pmatrix}\mathbf{v}_1^{\mathrm{T}}\\ \mathbf{v}_2^{\mathrm{T}}\\ \vdots\\ \mathbf{v}_n^{\mathrm{T}}\end{pmatrix} = \begin{pmatrix}(\mathbf{v}_1)_1 & (\mathbf{v}_1)_2 & \cdots & (\mathbf{v}_1)_n\\ (\mathbf{v}_2)_1 & (\mathbf{v}_2)_2 & \cdots & (\mathbf{v}_2)_n\\ \vdots & \vdots & \ddots & \vdots\\ (\mathbf{v}_n)_1 & (\mathbf{v}_n)_2 & \cdots & (\mathbf{v}_n)_n\end{pmatrix} \tag{5.31}$$

This matrix is not symmetric, but it is orthogonal, meaning that $\mathbf{\Gamma}^{\mathrm{T}}=\mathbf{\Gamma}^{-1}$. We can see this from

$$\begin{aligned}\mathbf{\Gamma}\mathbf{\Gamma}^{\mathrm{T}} &= \begin{pmatrix}\mathbf{v}_1^{\mathrm{T}}\\ \mathbf{v}_2^{\mathrm{T}}\\ \vdots\\ \mathbf{v}_n^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 & \cdots & \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_n\\ \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 & \cdots & \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_n\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{v}_n^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_n^{\mathrm{T}}\mathbf{v}_2 & \cdots & \mathbf{v}_n^{\mathrm{T}}\mathbf{v}_n\end{pmatrix} = \begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix} = \mathbf{1}\end{aligned} \tag{5.32}$$

This matrix $\mathbf{\Gamma}$ can be thought of a transformation from the original basis to the eigenbasis for $\mathbf{\Sigma}$. One effect of this is that it *diagonalizes* $\mathbf{\Sigma}$:

$$\mathbf{\Gamma}\mathbf{\Sigma}\mathbf{\Gamma}^{\mathrm{T}} = \begin{pmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{pmatrix} = \mathbf{\Lambda} \tag{5.33}$$

Finally, recall that the moment generating function is defined as

$$M(\mathbf{t}) = E(e^{\mathbf{t}\mathbf{X}}) \tag{5.34}$$

and that if we define $\psi(\mathbf{t}) = \ln M(\mathbf{t})$,

$$\left.\frac{\partial\psi}{\partial t_i}\right|_{\mathbf{t}=\mathbf{0}} = \mu_i \tag{5.35}$$

and

$$\left.\frac{\partial^2\psi}{\partial t_i\partial t_j}\right|_{\mathbf{t}=\mathbf{0}} = \operatorname{Cov}(X_1,X_2) \tag{5.36}$$

This means that if we do a Maclaurin expansion of $\psi(\mathbf{t})$ we get, in general,

$$\psi(\mathbf{t}) = \mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t} + \dots \tag{5.37}$$

where the terms indicated by $\dots$ have three or more powers of $\mathbf{t}$.

5.4 General Multivariate Normal Distribution

We define a multivariate normal random vector X as a random vector having the moment generating function

$$M(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t}\right) \tag{5.38}$$

We refer to the distribution as $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that this is equivalent to starting with the Maclaurin series for $\psi(\mathbf{t}) = \ln M(\mathbf{t})$ and cutting it off after the quadratic term.

We start with the mgf rather than the pdf because it applies whether the variance-covariance matrix $\boldsymbol{\Sigma}$ is positive definite or only positive semi-definite, i.e., whether it has one or more zero eigenvalues. To see what happens if one or more of the eigenvalues is zero, we use the orthonormal eigenvectors $\{\mathbf{v}_i\}$ of $\boldsymbol{\Sigma}$ to combine the random variables in $\mathbf{X}$ into n uncorrelated random variables $\{\mathcal{X}_i\}$, where $\mathcal{X}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{X}$, which have means $E(\mathcal{X}_i) = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}$ and variances $\operatorname{Var}(\mathcal{X}_i) = \lambda_i$. If we combine the $\{\mathcal{X}_i\}$ into a random vector

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\Gamma}\mathbf{X} \tag{5.39}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{pmatrix} \tag{5.40}$$

is the orthogonal matrix made up out of eigenvector components, $\boldsymbol{\mathcal{X}}$ has mean $E(\boldsymbol{\mathcal{X}}) = \boldsymbol{\Gamma}\boldsymbol{\mu}$ and variance-covariance matrix

$$\operatorname{Cov}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\Lambda} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{\mathrm{T}} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \tag{5.41}$$

The random vector $\boldsymbol{\mathcal{X}}$ also follows a multivariate normal distribution, in this case $N_n(\boldsymbol{\Gamma}\boldsymbol{\mu}, \boldsymbol{\Lambda})$. To show this, we'll show the more general result (which is Hogg's theorem 3.5.1), that if $\mathbf{X}$ is a $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ multivariate normal random vector, $\mathbf{A}$ is an $m \times n$ constant matrix and $\mathbf{b}$ is an m-element column vector, the random vector $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ also obeys a multivariate normal distribution. (Note that this works whether m is equal to, less than, or greater than n!) We prove this using the mgf. The mgf for $\mathbf{Y}$ is

$$\begin{aligned} M_Y(\mathbf{t}) &= E[\exp(\mathbf{t}^{\mathrm{T}}\mathbf{Y})] = E[\exp(\mathbf{t}^{\mathrm{T}}\mathbf{A}\mathbf{X} + \mathbf{t}^{\mathrm{T}}\mathbf{b})] \\ &= e^{\mathbf{t}^{\mathrm{T}}\mathbf{b}} E[\exp([\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\mathbf{X})] \end{aligned} \tag{5.42}$$

Now here is the key step, which Hogg doesn't elaborate on. $\mathbf{t}$ is an m-element column vector. $\mathbf{A}$ is an $m \times n$ matrix, so its transpose $\mathbf{A}^{\mathrm{T}}$ is an $n \times m$ matrix, and the combination $\mathbf{A}^{\mathrm{T}}\mathbf{t}$ is an n-element column vector, whose transpose is the n-element row vector

$$[\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}} = \mathbf{t}^{\mathrm{T}}\mathbf{A} \tag{5.43}$$

Therefore, the expectation value in the last line above is just the mgf for the original multivariate normal random vector $\mathbf{X}$

evaluated at the argument $\mathbf{A}^{\mathrm{T}}\mathbf{t}$:

$$\begin{aligned} E[\exp([\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\mathbf{X})] &= M_X(\mathbf{A}^{\mathrm{T}}\mathbf{t}) \\ &= \exp\left([\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}[\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\boldsymbol{\Sigma}[\mathbf{A}^{\mathrm{T}}\mathbf{t}]\right) \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\mathbf{t}\right) \end{aligned} \tag{5.44}$$

This makes the mgf for $\mathbf{Y}$ equal to $e^{\mathbf{t}^{\mathrm{T}}\mathbf{b}}$ times this, or

$$M_Y(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}] + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\mathbf{t}\right) \tag{5.45}$$

which is the mgf for a normal random vector with mean $\mathbf{A}\boldsymbol{\mu}+\mathbf{b}$ and variance-covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}$, i.e., one that obeys a $N_m(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}})$ distribution.

So, we return to the random vector $\boldsymbol{\mathcal{X}} = \boldsymbol{\Gamma}\mathbf{X}$, which we now see is a multivariate normal random vector with mean $\boldsymbol{\Gamma}\boldsymbol{\mu}$ and diagonal variance-covariance matrix $\boldsymbol{\Lambda}$. Its mgf is

$$\begin{aligned} M_{\boldsymbol{\mathcal{X}}}(\mathbf{t}) &= \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{t}\right) = \exp\left(\sum_{i=1}^{n}[t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\lambda_i t_i^2]\right) \\ &= \prod_{i=1}^{n}\exp\left(t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\lambda_i t_i^2\right) = \prod_{i=1}^{n} M_{\mathcal{X}_i}(t_i) \end{aligned} \tag{5.46}$$

which is the mgf of n independent random variables. There are two possibilities: either $\boldsymbol{\Sigma}$ (and thus $\boldsymbol{\Lambda}$) is positive definite, which means all of the $\{\lambda_i\}$ are positive, or one or more of the $\{\lambda_i\}$ are zero.

In the first case, we have the special case we considered before, n independent normally-distributed random variables, so the joint pdf is

$$f_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\xi}) = \prod_{i=1}^{n} f_{\mathcal{X}_i}(\xi_i) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Lambda})}}\exp\left(-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\Gamma}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\xi} - \boldsymbol{\Gamma}\boldsymbol{\mu})\right) \tag{5.47}$$

We can then do a multivariate transformation to get the pdf for $\mathbf{X} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\mathcal{X}} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\mathcal{X}}$. The Jacobian of the transformation is $\boldsymbol{\Gamma}^{\mathrm{T}}$, whose determinant is either 1 or -1, because

$$1 = \det \mathbf{1} = \det(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}) = (\det \boldsymbol{\Gamma})^2 \tag{5.48}$$

(This is true for any orthogonal matrix.) This means $|\det \boldsymbol{\Gamma}| = 1$, and the pdf for $\mathbf{X}$ is simply

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\Gamma}\mathbf{x}) \tag{5.49}$$

If we also note that $\det \boldsymbol{\Lambda} = \prod_{i=1}^{n}\lambda_i = \det \boldsymbol{\Sigma}$, we see that

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Lambda})}}\exp\left(-\frac{1}{2}\boldsymbol{\Gamma}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\mathbf{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{5.50}$$

But the combination $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}$ is just the inverse of $\boldsymbol{\Sigma}$, because

$$(\boldsymbol{\Lambda})^{-1} = (\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{\mathrm{T}})^{-1} = (\boldsymbol{\Gamma}^{\mathrm{T}})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}^{-1} \tag{5.51}$$

so we find that, for arbitrary positive definite symmetric $\boldsymbol{\Sigma}$,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{5.52}$$

which is exactly the generalization we expected from the n-independent-random-variable case.

On the other hand, if $\boldsymbol{\Sigma}$ has one or more zero eigenvalues, so that $\det\boldsymbol{\Sigma} = 0$, and $\boldsymbol{\Sigma}^{-1}$ is not defined, that pdf won't make sense. In that case, consider the random variable $\mathcal{X}_i$ corresponding to the zero eigenvalue $\lambda_i = 0$. Its mgf is

$$E(e^{t_i\mathcal{X}_i}) = M_{\mathcal{X}_i}(t_i) = \exp\left(t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\lambda_i t_i^2\right) = \exp\left(t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}\right) \tag{5.53}$$

but the only way that is possible is if $\mathcal{X}_i$ is always equal to $\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}$, i.e., $\mathcal{X}_i$ is actually a discrete random variable with pmf

$$P(\mathcal{X}_i = \xi_i) = \begin{cases} 1 & \xi_i = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} \\ 0 & \text{otherwise} \end{cases} \tag{5.54}$$

This is the limit of a normal distribution as its variance goes to zero.

Returning to the case where $\boldsymbol{\Sigma}$ is positive definite, so that each $\mathcal{X}_i$ is an independent $N(\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}, \lambda_i)$ random variable, we can construct the corresponding standard normal random variable

$$\mathcal{Z}_i = (\lambda_i)^{-1/2}(\mathcal{X}_i - \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}) = (\lambda_i)^{-1/2}\mathbf{v}_i^{\mathrm{T}}(X_i - \boldsymbol{\mu}) \tag{5.55}$$

which we could combine into a $N_n(\mathbf{0},\mathbf{1})$ random vector

$$\boldsymbol{\mathcal{Z}} = \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \\ \vdots \\ \mathcal{Z}_n \end{pmatrix} \tag{5.56}$$

However, it's actually more convenient to combine them into a different $N_n(\mathbf{0},\mathbf{1})$ random vector

$$\boldsymbol{Z} = \sum_{i=1}^{n} \mathbf{v}_i\mathcal{Z}_i = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\mathcal{Z}} = \sum_{i=1}^{n} \mathbf{v}_i(\lambda_i)^{-1/2}\mathbf{v}_i^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{1/2}(\mathbf{X}-\boldsymbol{\mu}) \tag{5.57}$$

with pdf

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}}e^{-\mathbf{z}^{\mathrm{T}}\mathbf{z}/2} \tag{5.58}$$

Hogg uses this as the starting point for deriving the pdf for the multivariate normal random vector $\mathbf{X}$, with the factor of

$$\frac{1}{\sqrt{\det\boldsymbol{\Sigma}}} = \left|\det(\boldsymbol{\Sigma}^{-1/2})\right| \tag{5.59}$$

coming from the Jacobian determinant associated with the transformation.

Tuesday 29 October 2013 – Read Section 3.6 of Hogg

5.5 Consequences of the Multivariate Normal Distribution

Recall that if $\mathbf{X}$ is a multivariate normal random vector, obeying a $N_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$ distribution, this is equivalent to saying $\mathbf{X}$ has the its mgf

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}\mathbf{X} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t}\right) \tag{5.60}$$

Its mean is $\mathbf{X} = \boldsymbol{\mu}$ and its variance-covariance matrix is $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$. If the variance-covariance matrix is invertible, then $\mathbf{X}$ has the probability density function

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right) \tag{5.61}$$

5.5.1 Connection to χ^2 Distribution

We also showed that we could make a random vector

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \boldsymbol{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \tag{5.62}$$

which obeyed a $N_n(\mathbf{0}, \mathbf{1})$ distribution, i.e., where the $\{Z_i\}$ were independent standard normal random variables. Previously, we showed that if we took the sums of the squares of n independent standard normal random variables, the resulting random variable obeyed a chi-square distribution with n degrees of freedom. So in this case we can construct

$$\begin{aligned} Y = \sum_{i=1}^{n} (Z_i)^2 = \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} &= [\mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})]^{\mathrm{T}} [\mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})] \\ &= (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1/2} \mathbf{\Sigma}^{-1/2} (\mathbf{X} - \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \end{aligned} \tag{5.63}$$

and it will be a $\chi^2(n)$ random variable. This is Theorem 3.5.4 of Hogg. Note that if Σ is diagonal, so that $\{X_i\}$ are n independent random variables, with X_i being $N(\mu_i, \sigma_i^2)$, the chi-square random variable reduces to

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \qquad \text{if } \mathbf{\Sigma} \text{ diagonal} \tag{5.64}$$

5.6 Marginal Distribution

One last question to consider is, what if we split the n random variables in the random vector $\mathbf{X}$ into two groups: the first m, which we collect into a random vector $\mathbf{X}_1$, and the last $p = m - n$, which we collect into a random vector $\mathbf{X}_2$, so that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \tag{5.65}$$

where

$$\mathbf{X}_1 = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \qquad \text{and} \qquad \mathbf{X}_2 = \begin{pmatrix} X_{m+1} \\ X_{m+2} \\ \vdots \\ X_n \end{pmatrix} \tag{5.66}$$

We'd like to know what the marginal distributions for $\mathbf{X}_1$ and $\mathbf{X}_2$ are, and also the conditional distributions for $\mathbf{X}_1|\mathbf{X}_2$ and $\mathbf{X}_2|\mathbf{X}_1$. As a bit of bookkeeping, we partition the mean vector

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \tag{5.67}$$

and the variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \tag{5.68}$$

in the same way. $\boldsymbol{\mu}_1$ is an m-element column vector, $\boldsymbol{\mu}_2$ is a p-element column vector, $\mathbf{\Sigma}_{11}$ is a $m \times m$ symmetric matrix, $\mathbf{\Sigma}_{22}$ is a $p \times p$ symmetric matrix, $\mathbf{\Sigma}_{12}$ is a $m \times p$ matrix (i.e., m rows and p columns), and $\mathbf{\Sigma}_{21}$ is a $p \times m$ matrix. Since $\mathbf{\Sigma}$ is symmetric, we must have

$$\mathbf{\Sigma} = \mathbf{\Sigma}^{\mathrm{T}} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{21}^{\mathrm{T}} \\ \mathbf{\Sigma}_{12}^{\mathrm{T}} & \mathbf{\Sigma}_{22} \end{pmatrix} \tag{5.69}$$

so $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{12}^{\mathrm{T}}$. Now, it's pretty easy to get the marginal distributions using the mgf. Partitioning $\mathbf{t}$ in the usual way, we see

$$\mathbf{t}^{\mathrm{T}} \boldsymbol{\mu} = \mathbf{t}_1^{\mathrm{T}} \boldsymbol{\mu}_1 + \mathbf{t}_2^{\mathrm{T}} \boldsymbol{\mu}_2 \tag{5.70}$$

and

$$\begin{aligned}\mathbf{t}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{t} &= \begin{pmatrix}\mathbf{t}_1^{\mathrm{T}} & \mathbf{t}_2^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{t}_1\\ \mathbf{t}_2\end{pmatrix}\\ &= \mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{t}_1 + \mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{12}\mathbf{t}_2 + \mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{21}\mathbf{t}_1 + \mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{22}\mathbf{t}_2\\ &= \mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{t}_1 + 2\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{12}\mathbf{t}_2 + \mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{22}\mathbf{t}_2\end{aligned} \tag{5.71}$$

where in the last step we've used the fact that since $\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{12}\mathbf{t}_2$ is a 1×1 matrix, i.e., a number, it is its own transpose:

$$\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{12}\mathbf{t}_2 = (\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{12}\mathbf{t}_2)^{\mathrm{T}} = \mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{12}^{\mathrm{T}}\mathbf{t}_1 = \mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{21}\mathbf{t}_1 \tag{5.72}$$

Thus the joint mgf is

$$\begin{aligned}M(\mathbf{t}) &= M(\mathbf{t}_1, \mathbf{t}_2)\\ &= \exp\left(\mathbf{t}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \mathbf{t}_2^{\mathrm{T}}\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{t}_1 + \mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{12}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{22}\mathbf{t}_2\right)\end{aligned} \tag{5.73}$$

The mgf for $\mathbf{X}_1$ is thus

$$M_1(\mathbf{t}_1) = M(\mathbf{t}_1, \mathbf{0}) = \exp\left(\mathbf{t}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \frac{1}{2}\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{t}_1\right) \tag{5.74}$$

which means $\mathbf{X}_1$ is a $N_m(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$ multivariate normal random vector, and likewise the mgf for $\mathbf{X}_2$ is thus

$$M_2(\mathbf{t}_2) = M(\mathbf{0}, \mathbf{t}_2) = \exp\left(\mathbf{t}_2^{\mathrm{T}}\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{22}\mathbf{t}_2\right) \tag{5.75}$$

so $\mathbf{X}_2$ is $N_p(\boldsymbol{\mu}_2, \mathbf{\Sigma}_{22})$. (Hogg also shows this as a special case of the more general result that $\mathbf{A}\mathbf{X} + \mathbf{b}$ is a normal random vector, where $\mathbf{A}$ is a $m \times n$ constant matrix and $\mathbf{b}$ is a m-element constant column vector.) We see that

$$M_1(\mathbf{t}_1)M_2(\mathbf{t}_2) = \exp\left(\mathbf{t}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \mathbf{t}_2^{\mathrm{T}}\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{t}_1 + \frac{1}{2}\mathbf{t}_2^{\mathrm{T}}\mathbf{\Sigma}_{22}\mathbf{t}_2\right) \tag{5.76}$$

which is only equal to $M(\mathbf{t}_1, \mathbf{t}_2)$ if the cross-part $\mathbf{\Sigma}_{12}$ of the variance-covariance matrix is zero. So $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\mathbf{\Sigma}_{12} = \mathbf{0}_{m \times p}$.

5.7 Conditional Distribution

Finally, we'd like to consider the conditional distribution of $\mathbf{X}_1$ given that $\mathbf{X}_2 = \mathbf{x}_2$. Hogg proves that the distribution is a multivariate normal, specifically $N_m(\boldsymbol{\mu}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}[\mathbf{x}_2 - \boldsymbol{\mu}_2], \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21})$, by using the mgf, but their proof assumes you already know the form of the distribution. The proof assumes that $\mathbf{\Sigma}$ is positive definite, and in particular that $\mathbf{\Sigma}_{22}^{-1}$ exists. We might like to try to work this out rather than starting with the answer, and so we could divide the pdf

$$f(\mathbf{x}) = f(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right) \tag{5.77}$$

by the marginal pdf

$$f_2(\mathbf{x}_2) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma}_{22})}}\exp\left(-\frac{1}{2}(\mathbf{X}_2 - \boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)\right) \tag{5.78}$$

to get the conditional pdf

$$f_{1|2}(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f_2(\mathbf{x}_2)} \tag{5.79}$$

Working out the details of this is a little nasty, though, since it requires a block decomposition of $\mathbf{\Sigma}^{-1}$, which exists, but is kind of complicated. (For example $[\mathbf{\Sigma}^{-1}]_{11} = [\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}]^{-1}$.) It's not hard to see that what comes out will a Gaussian in $\mathbf{x}_1$ minus something, though. Still, for simplicity we can limit the explicit demonstration to the case where $m = 1$ and $p = 1$, so

$n = 2$. Then all of the "blocks" in the decomposition of the $n \times n$ matrices are just numbers, specifically

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{5.80}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) \\ \operatorname{Cov}(X_1, X_2) & \operatorname{Var}(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{5.81}$$

familiar results from matrix algebra tell us that

$$\det \boldsymbol{\Sigma} = \begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2(1-\rho^2) \tag{5.82}$$

and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \tag{5.83}$$

We know that the correlation coefficient ρ obeys $\rho^2 \leq 1$; in order for $\boldsymbol{\Sigma}$ to be positive definite, we require $\rho^2 < 1$. You will show on the homework that the joint pdf for this bivariate normal distribution is

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2(1-\rho^2)} + \frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2(1-\rho^2)}\right) \end{aligned} \tag{5.84}$$

and since the marginal distribution for X_2 is

$$f_2(x_2) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right) \tag{5.85}$$

the conditional pdf will be

$$\begin{aligned} f_{1|2}(x_1|x_2) &= \frac{f(x_1,x_2)}{f_2(x_2)} = \frac{1}{\sigma_1\sqrt{(1-\rho^2)2\pi}} \\ &\times \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2(1-\rho^2)} + \frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} - \frac{\rho^2(x_2-\mu_2)^2}{2\sigma_2^2(1-\rho^2)}\right) \\ &= \frac{1}{\sigma_1\sqrt{(1-\rho^2)2\pi}} \exp\left(-\frac{1}{2\sigma_1^2(1-\rho^2)}\left[x_1 - \mu_1 - \frac{\rho\sigma_1}{\sigma_2}(x_2-\mu_2)\right]^2\right) \end{aligned} \tag{5.86}$$

which is a normal distribution with mean $\mu_1 - \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and variance $\sigma_1^2(1-\rho^2)$.

Thursday 31 October 2013 – Read Section 4.1 of Hogg

6 The t- and F Distributions

Our final distributions are two distributions related to the normal and chi-square distributions, which are very useful in statistical inference.

6.1 The t distribution

If Z is a standard normal random variable $[N(0,1)]$ and V is a chi-square random variable with r degrees of freedom $[\chi^2(r)]$, and Z and V are independent, the combination

$$T = \frac{Z}{\sqrt{V/r}} \tag{6.1}$$

is a random variable following a t-distribution with r degrees of freedom; its pdf is

$$f_T(t) \propto \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}} \qquad -\infty < t < \infty \tag{6.2}$$

where we have not written the explicit normalization constant, in the interest of notational simplicity. This pdf can be derived by the change of variables formalism. If r is very large, the t-distribution is approximately the same as the standard normal distribution. As you'll show on the homework, if $r = 1$, it's the Cauchy distribution. Note that the moment generating function for the t-distribution doesn't exist, since for $n \geq r$, the integral

$$\int_{-\infty}^{\infty} t^n \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}} dt \tag{6.3}$$

diverges, as its integrand becomes, up to a constant, t^{n-r-1} for large t.

6.2 The F distribution

If U is a $\chi^2(r_1)$ random variable and V is a $\chi^2(r_2)$ random variable, and U and V are independent, the combination

$$F = \frac{U/r_1}{V/r_2} \tag{6.4}$$

obeys an F distribution. Again its pdf can be worked out by the change of variables method, and is, up to the normalization constant

$$f_F(x) \propto x^{\frac{r_1}{2}-1} \left(1 + \frac{r_1}{r_2} x\right)^{-\frac{r_1+r_2}{2}} \qquad 0 < x < \infty \tag{6.5}$$

6.3 Student's Theorem

We can illustrate the usefulness of the t-distribution, also known as Student's t-distribution, by considering a result by William S. Gosset, writing under the pseudonym of "Student" while working in the Guinness brewery in Dublin. Suppose we have n iid normal random variables, each with mean μ and variance σ^2, i.e., $\mathbf{X}$ is a random vector with mean

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu\,\mathbf{e} \tag{6.6}$$

where we have defined the n-element column vector

$$\mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{6.7}$$

and variance-covariance matrix

$$\operatorname{Cov}(\mathbf{X}) = \sigma^2 \mathbf{1}_{n\times n} \tag{6.8}$$

We know from elementary statistics that the sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{6.9}$$

and sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{6.10}$$

can be used to estimate the mean μ and variance σ^2, respectively, i.e., $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$, and also that $\operatorname{Var}(\overline{X}) = \sigma^2/n$. Student's theorem says that the following things are also true:

1. $\overline{X}$ obeys a normal distribution, i.e., it is $N(\mu, \sigma^2/n)$;

2. $\overline{X}$ and S^2 are independent;

3. The combination
$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \tag{6.11}$$
is a chi-square random variable with $n-1$ degrees of freedom $[\chi^2(n-1)]$;

4. The combination
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{6.12}$$
obeys a t-distribution with $n-1$ degrees of freedom.

This theorem underlies all of the confidence intervals you've ever constructed for the mean of a normal distribution with unknown variance using a small random sample drawn from that distribution.

The demonstration that Student's theorem is true is a nice application of the multivariate normal distribution. First, we can write
$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \mathbf{e}^{\mathrm{T}} \mathbf{X} \tag{6.13}$$
and
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n-1} (\mathbf{X} - \mathbf{e}\overline{X})^{\mathrm{T}} (\mathbf{X} - \mathbf{e}\overline{X}) \tag{6.14}$$
Note that $\overline{X} = \frac{1}{n}\mathbf{e}^{\mathrm{T}}\mathbf{X}$ and
$$\mathbf{Y} = \mathbf{X} - \mathbf{e}\overline{X} = \left(\mathbf{1} - \frac{1}{n} \mathbf{e}\mathbf{e}^{\mathrm{T}} \right) \mathbf{X} \tag{6.15}$$
are both linear transformations of $\mathbf{X}$. We can combine them into an $n+1$-element random vector
$$\begin{pmatrix} \overline{X} \\ X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{6.16}$$
$$= \begin{pmatrix} \frac{1}{n}\mathbf{e}^{\mathrm{T}} \\ \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \end{pmatrix} \mathbf{X} = \mathbf{A}\mathbf{X}$$
where $\mathbf{A}$ is a $(n+1)\times n$ matrix. Since the original random vector $\mathbf{X}$ has $\boldsymbol{\mu} = \mu\mathbf{e}$ and, the transformed vector $\mathbf{A}\mathbf{X}$ has expectation value
$$E(\mathbf{A}\mathbf{X}) = \mathbf{A}\boldsymbol{\mu} = \begin{pmatrix} \frac{1}{n}\mathbf{e}^{\mathrm{T}} \\ \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \end{pmatrix} \mu\mathbf{e} = \begin{pmatrix} \mu \\ \mathbf{0} \end{pmatrix} \tag{6.17}$$
where we've used the fact that
$$\mathbf{e}^{\mathrm{T}}\mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} = n \tag{6.18}$$
and
$$(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}})\mathbf{e} = \mathbf{e} - \frac{n}{n}\mathbf{e} = \mathbf{0} \tag{6.19}$$

Since $\mathbf{X}$ has the variance-covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{1}$, the transformed random vector $\mathbf{AX}$ has variance-covariance matrix[4]

$$\begin{aligned}\operatorname{Cov}(\mathbf{AX}) = \mathbf{A\Sigma A}^{\mathrm{T}} &= \begin{pmatrix} \frac{1}{n}\mathbf{e}^{\mathrm{T}} \\ \mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}} \end{pmatrix} \sigma^2\mathbf{1} \begin{pmatrix} \frac{1}{n}\mathbf{e} & \mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}} \end{pmatrix} \\ = \sigma^2 \begin{pmatrix} \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{n^2} & \mathbf{e}^{\mathrm{T}}\left(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}\right) \\ \left(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}\right)\mathbf{e} & \left(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}\right)^2 \end{pmatrix} &= \sigma^2 \begin{pmatrix} \frac{1}{n} & \mathbf{0}_{1\times n} \\ \mathbf{0}_{n\times 1} & \left(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}\right) \end{pmatrix}\end{aligned} \tag{6.20}$$

Where we've used the fact (previously noted) that $\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}$ annihilated $\mathbf{e}$ and also that

$$(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}})^2 = \mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}} + \frac{1}{n^2}\mathbf{e}n\mathbf{e}^{\mathrm{T}} = \mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}} \tag{6.21}$$

We say that a matrix with this property is a *projection matrix*, in this case onto n-dimensional vectors perpendicular to $\mathbf{e}$.

Since the first row and column of $\operatorname{Cov}(\mathbf{AX})$ are all zeros, except for the diagonal element, it means that the random variable $\overline{X}$, which is the first element of $\mathbf{AX}$, and the random vector $\mathbf{Y}$ are independent, which in turn means $\overline{X}$ and $S^2 = \frac{1}{n-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ are independent random variables, which is part 2 of Student's theorem. We've also seen part 1, since $\overline{X}$ is a normal random variable whose mean is the first element of $E[\mathbf{AX}]$, i.e., μ and whose variance is the $(1,1)$ element of $\operatorname{Cov}(\mathbf{AX})$, which is σ^2/n.

To see that

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{Y}}{\sigma^2} \tag{6.22}$$

is a chi-square random variable with $n-1$ degrees of freedom, consider the variance-covariance matrix

$$\operatorname{Cov}(\mathbf{Y}) = \left(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}\right)\sigma^2 \tag{6.23}$$

This is an $n\times n$ matrix, but it is not invertable, so we can't do the usual trick to construct a $\chi^2(n)$ random variable. It's actually not too hard to work out the eigenvalue decomposition of this matrix; since $(\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}})\mathbf{e} = \mathbf{0}$, we see that $\mathbf{e}$ is an eigenvector with eigenvalue zero. Because the matrix $\mathbf{1} - \frac{1}{n}\mathbf{ee}^{\mathrm{T}}$ is a projector onto the $n-1$-dimensional subspace perpendicular to $\mathbf{e}$, we can choose any $n-1$ orthonormal vectors in that subspace, and they will be eigenvectors with eigenvalue σ^2. For example, take

$$\mathbf{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ -1\sqrt{2} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -2\sqrt{6} \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \cdots \\ \mathbf{v}_{n-1} = \begin{pmatrix} 1/\sqrt{n(n-1)} \\ 1/\sqrt{n(n-1)} \\ 1/\sqrt{n(n-1)} \\ \vdots \\ 1/\sqrt{n(n-1)} \\ -(n-1)/\sqrt{n(n-1)} \end{pmatrix} \tag{6.24}$$

if we let $\mathbf{v}_n = \mathbf{e}/\sqrt{n}$, we have our complete set of orthonormal eigenvectors, with $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = \sigma^2$ and $\lambda_n = 0$. If we define $\mathcal{Y}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{Y}$, then $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_{n-1}$ are $n-1$ independent normal random variables each with variance σ^2. (The last combination is trivial, since $\mathbf{v}_n^{\mathrm{T}}\mathbf{Y} = \frac{1}{\sqrt{n}}\mathbf{e}^{\mathrm{T}}\mathbf{Y} = \mathbf{0}$.) This means the

[4] Remember that $\mathbf{A}$ is $(n+1)\times n$, $\mathbf{\Sigma}$ is $n\times n$, and $\mathbf{A}^{\mathrm{T}}$ is $n\times(n-1)$, so $\mathbf{A\Sigma A}^{\mathrm{T}}$ is $(n+1)\times(n+1)$.

following combination is a $\chi^2(n-1)$ random variable:

$$\sum_{i=1}^{n-1}\left(\frac{\mathcal{Y}_i}{\sigma}\right)=\frac{\mathbf{Y}^{\mathrm{T}}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\mathbf{Y}}{\sigma^2}=\frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\left(\sum_{i=1}^{n-1}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right)\mathbf{Y}$$
$$=\frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\left(\mathbf{1}-\frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right)\mathbf{Y}=\frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}=\frac{n-1}{\sigma^2}S^2 \tag{6.25}$$

This is point 3 of Student's theorem.

Finally, for point 4 we construct a t-distributed random variable.

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \tag{6.26}$$

is a standard normal random variable, and

$$\frac{n-1}{\sigma^2}S^2 \tag{6.27}$$

is a $\chi^2(n-1)$, we can take

$$T=\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{n-1}{\sigma^2}S^2/(n-1)}}=\frac{\overline{X}-\mu}{\sqrt{S^2/n}} \tag{6.28}$$

and it will obey a t-distribution with $n-1$ degrees of freedom, which completes the proof of Student's theorem.