12

Appendix A: Background

The purpose of this Appendix is to review background material on the normal distribution and its relatives, and an outline of the basics of estimation and hypothesis testing as they are applied to problems arising from the normal distribution. Proofs are not given since it is assumed that the reader is familiar with the material from more elementary courses.

12.1 The Normal, Chi-squared, t and F distributions

A random variable Y is said to have a *normal distribution* with mean μ and variance σ^2 (notation: $Y \sim N(\mu, \sigma^2)$) if it is a continuous real-valued random variable with density

$$f(y; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-(y-\mu)^2/2\sigma^2\}. \tag{12.1}$$

The case where $\mu = 0$ and $\sigma^2 = 1$ is called *standard normal*.

Proposition 12.1. If Z is standard normal and $Y = \mu + \sigma Z$ then $Y \sim N(\mu, \sigma^2)$.

This result is particularly useful in calculating probabilities for a general normal random variable. The distribution function for standard normal, given by

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

cannot be evaluated analytically but is widely tabulated. To compute the distribution function for $Y \sim N(\mu, \sigma^2)$, define $Z = (Y - \mu)/\sigma$ and use Proposition 12.1 in the form

$$\Pr\{Y \le y\} = \Pr\left\{\frac{Y-\mu}{\sigma} \le \frac{y-\mu}{\sigma}\right\} = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

Other useful properties of the normal distribution are summarized in

Proposition 12.2. If $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$,..., $Y_n \sim N(\mu_n, \sigma_n^2)$ are *independent* normal random variables, then

$$\sum_{i=1}^{n} Y_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Note that the important statement here is that the sum has a normal distribution. The mean and variance follow from elementary calculations.

More generally, if $Y_1, ..., Y_n$ have any joint distribution with means $\mu_1, ..., \mu_n$ and covariances $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$, and if $a_1, ..., a_n$ are constants, then $\sum a_i Y_i$ has mean $\sum a_i \mu_i$ and variance $\sum \sum a_i a_j \sigma_{ij}$. If $Y_1, ..., Y_n$ are jointly normal then the sum has a normal distribution as well. The latter property (that all linear combinations of a set random variables have a normal distribution) may be taken as the definition of jointly normal.

As already mentioned, the standard normal distribution function Φ is widely tabulated. For calculations used in constructing hypothesis tests and confidence intervals, we often need to know the inverse standard normal distribution function, i.e. for given A we need to know z such that

$$\Phi(z) = A.$$

The resulting z is denoted z_A. Sometimes tables of z_A are produced; if they are not available, then it is necessary to interpolate in a table of Φ.

A random variable X has a *chi-squared distribution with n degrees of freedom* (notation: $X \sim \chi_n^2$) if it can be written in the form $X = Z_1^2 + Z_2^2 + ... + Z_n^2$ where $Z_1, ..., Z_n$ are independent standard normal. It is possible also to define the chi-squared distribution in terms of its density, but we shall not need that.

The most important property of the chi-squared distribution is that it is the sampling distribution of the sample variance in the case of normally distributed samples. Suppose $Y_1, ..., Y_n$ are a sample of observations. Define the sample mean and variance

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \tag{12.2}$$

Note in particular the divisor $n - 1$ (rather than n) in the definition of s^2. This is to make the estimator unbiased.

We then have:

Proposition 12.3. Suppose $Y_1, ..., Y_n$ are independent with distribution $N(\mu, \sigma^2)$. Then

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Moreover, $\bar{Y}$ and s^2 are independent random variables.

As with the normal distribution, it is useful to have some notation for the inverse of the chi-squared distribution function. Accordingly, we define the number $\chi^2_{n;A}$ by the following property:

$$\text{If } X \sim \chi^2_n \text{ then } \Pr\{X \leq \chi^2_{n;A}\} = A.$$

Statistical tables typically give values of either $\chi^2_{n;A}$ or $\chi^2_{n;1-A}$ for $n = 1, 2, ...,$ and a range of values of A.

The *t distribution with n degrees of freedom* is defined as the distribution of $T = Z/\sqrt{(X/n)}$ when Z and X are *independent*, $Z \sim N(0,1)$ and $X \sim \chi^2_n$. This is usually written $T \sim t_n$. The most important application is the following:

Proposition 12.4. If $Y_1, ..., Y_n$ are independent with distribution $N(\mu, \sigma^2)$, then

$$\sqrt{n}\frac{\bar{Y} - \mu}{s} \sim t_{n-1}.$$

Note that this follows at once from Proposition 12.3 and the definition of the t distribution.

The inverse distribution function is defined as the function $t_{n;A}$ with the property:

$$\text{If } T \sim t_n \text{ then } \Pr\{T \leq t_{n;A}\} = A.$$

Once again, this or some variant of it is tabulated in all sets of statistical tables.

The final distribution in this class is the *F distribution*, which is defined as follows. Let X_1 and X_2 be two independent chi-squared random variables with n_1 and n_2 degrees of freedom respectively. Let

$$U = \frac{X_1}{n_1} . \frac{n_2}{X_2}.$$

Then U has an *F distribution with n_1 and n_2 degrees of freedom* (notation: $U \sim F_{n_1,n_2}$). The inverse distribution function is denoted by $F_{n_1,n_2;A}$ with the property

$$\text{If } U \sim F_{n_1,n_2} \text{ then } \Pr\{U \leq F_{n_1,n_2;A}\} = A.$$

Again, this is tabulated in all sets of statistical tables. Sometimes it is necessary to use the identity

$$F_{n_1,n_2;A} = 1/F_{n_2,n_1;1-A}$$

which follows immediately from the definition of the F distribution.

The best known application of the F distribution is in comparing the variances of two samples. Suppose $Y_1, ..., Y_m$ are independent observations from $N(\mu_1, \sigma_1^2)$ and $W_1, ..., W_n$ an independent sample of independent observations from $N(\mu_2, \sigma_2^2)$. Suppose we are interested in the ratio σ_1^2/σ_2^2; for instance, we might want to test the hypothesis that this ratio is 1. Calculate the sample variances s_1^2 and s_2^2; then

$$U = \frac{s_1^2}{\sigma_1^2} \cdot \frac{\sigma_2^2}{s_2^2} \tag{12.3}$$

has an $F_{m-1,n-1}$ distribution. In particular, under the null hypothesis that $\sigma_1^2 = \sigma_2^2$ this reduces to the statement that s_1^2/s_2^2 has an F distribution.

12.2 Estimation and hypothesis testing: The normal means problem

To begin our review of estimation and hypothesis testing, we shall discuss the problem of estimating μ when $Y_1, ..., Y_n$ are independent from $N(\mu, \sigma^2)$ and σ^2 is known. In most contexts it is unrealistic to assume that σ^2 is known while μ is unknown, and the case where both parameters are unknown is considered in the next section. There are some situations, such as trying to "tune" the mean level of a piece of machinery which has already been operating long enough for the variance to be assumed known, in which the present formulation may be realistic. However, the main reason for considering the present problem first is that it is the simplest of its type, and therefore serves to define a framework which will be useful in studying other problems later on.

The natural estimator of μ is the sample mean $\bar{Y}$. This has a number of desirable properties; for example, it is unbiased and is a minimum variance unbiased estimator. It is also the maximum likelihood estimator of μ. In view of Proposition 12.3 we may define

$$Z = \sqrt{n}\frac{\bar{Y} - \mu}{\sigma} \tag{12.4}$$

which then has a standard normal distribution.

Suppose we want to form a $100(1-\alpha)\%$ confidence interval for μ, where $0 < \alpha < 1$ is given. Consider the following sequence of equalities:

$$\begin{aligned}1-\alpha &= \Pr\{-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}\} \\ &= \Pr\left\{\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{Y} \le \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} \\ &= \Pr\left\{\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}. \qquad (12.5)\end{aligned}$$

The last inequality has μ in the middle, and is therefore in the form we need to specify a confidence interval. The conclusion is that the interval

$$\left[\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad (12.6)$$

is the desired confidence interval. The interpretation of this interval is that, in a long run of experiments conducted under identical conditions, the quoted interval (12.6) will include the true mean μ a proportion $1-\alpha$ of the time, provided of course all the assumptions that have been made are correct.

For example, in applications it is quite common to take $\alpha = 0.05$, corresponding to which $z_{0.975} = 1.96$. The interval

$$\left[\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

Now let us turn to hypothesis testing. We consider the following three possible specifications of the null hypothesis H_0 and the alternative H_1:

Problem A. $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$.

Problem B. $H_0 : \mu \ge \mu_0$ versus $H_1 : \mu < \mu_0$.

Problem C. $H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$.

In each of these, μ_0 is a specified numerical value. The three possibilities A,B and C are certainly not the only ones it is possible to consider, but they cover the majority of practical situations.

In each case, the test procedure consists of first forming a *test statistic* which summarizes the information in the sample about the unknown parameter μ, and then forming a *rejection region* which determines the values of the test statistic for which the null hypothesis H_0 is rejected. The form of the rejection region depends on the structure of the null and alternative hypotheses.

For this problem the natural test statistic is $\bar{Y}$, and the form of rejection region depends on which of the above three testing problems we are considering:

For problem A: reject H_0 if $\bar{Y} > c_A$.

For problem B: reject H_0 if $\bar{Y} < c_B$.

For problem C: reject H_0 if $|\bar{Y} - \mu| > c_C$.

In each case the constant c_A, c_B or c_C is chosen to satisfy the probability requirement that the probability of rejection of the null hypothesis, when the null hypothesis is true, should not be less than α, where $0 < \alpha < 1$ is specified.

In the case of problem A, suppose first we take $\mu = \mu_0$. Defining Z as in (12.4) with $\mu = \mu_0$, we quickly deduce

$$\begin{aligned}\alpha &= \Pr\{Z > z_{1-\alpha}\}\\ &= \Pr\left\{\bar{Y} > \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\}\end{aligned}$$

from which we deduce that we should take $c_A = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. It is then readily checked that, for any other μ in H_0, i.e. for $\mu < \mu_0$, the probability that $\bar{Y} > c_A$ is smaller than α, so the probability requirement is satisfied.

As an example, if we take $\alpha = 0.05$ then $z_{0.95} = 1.645$ so the appropriate test is to reject H_0 if $\bar{Y}$ is bigger than $\mu + 1.645\sigma/\sqrt{n}$.

The argument for problem B is exactly similar, but with all the signs reversed. We reject H_0 if $\bar{Y} < c_B$, where $c_B = \mu_0 - \sigma z_{1-\alpha}/\sqrt{n}$.

In the case of problem C, the same sequence of inequalities as in (12.5), with $\mu = \mu_0$, leads us to deduce that we should take $c_C = z_{1-\alpha/2}\sigma/\sqrt{n}$. Note that, as in (12.5) but in contrast to the results for problems A and B, the appropriate point of the normal distribution is now $z_{1-\alpha/2}$, not $z_{1-\alpha}$. The difference reflects the fact that we are dealing with a *two-sided* (alternative) hypothesis, whereas in cases A and B the alternative hypotheses are both one-sided.

12.3 Other common estimation and testing problems

In this section we consider a number of other standard problems.

12.3.1 Normal means, variance unknown

Suppose we have a sample $Y_1, ..., Y_n$ of independent observations from $N(\mu, \sigma^2)$, but this time with both μ and σ^2 unknown. Again, our interest is in forming a confidence interval or testing a hypothesis about μ. The key difference is that we replace equation (12.4) by

$$t = \sqrt{n}\frac{\bar{Y} - \mu}{s} \qquad (12.7)$$

where s is the sample standard deviation. Then t has the distribution of t_{n-1}. All the results of Section 12.2 remain valid, except that wherever σ

appears we replace it by its estimate s, and wherever a normal distribution point z_A appears we replace it by its corresponding value $t_{n-1;A}$. Thus, for example, a $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\bar{Y} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{Y} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]; \tag{12.8}$$

compare with equation (12.6). For example, with $n = 5, 10$ and 20, the respective t values for $\alpha = 0.05$ are 2.776, 2.262 and 2.093 for 4, 9 and 19 degrees of freedom, compared with the limiting value 1.96 for the normal distribution. This indicates the extent to which the confidence interval must be lengthened to allow for the estimation of σ; even for $n = 10$ the effect is quite modest, resulting in a 15% $(2.262/1.96 = 1.15)$ lengthening of the confidence interval.

The quantity $s/\sqrt{n}$ is known as the *standard error* of the estimate $\bar{Y}$; it represents our estimate of the standard deviation of $\bar{Y}$, after substituting the estimate s for the unknown true residual standard deviation σ.

When testing the null hypothesis $\mu = 0$, the statistic t reduces to $\sqrt{n}\bar{Y}/s$; in other words, the sample estimate of μ, divided by its standard error. Very generally in statistics, when we take an estimate of a parameter and divide it by its standard error, we call the resulting quantity the *t statistic*. It forms the basis for very many tests of hypotheses.

12.3.2 Comparison of two normal means, variances known

Suppose now we have two samples, $Y_1, ..., Y_m$ and $W_1, ..., W_n$, respectively from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with all observations independent and σ_1^2 and σ_2^2 known. Our interest is in tests or confidence intervals for the difference of means, $\mu_1 - \mu_2$. Of particular interest is the possibility of testing whether $\mu_1 = \mu_2$, or in other words whether $\mu_1 - \mu_2 = 0$, against either one-sided or two-sided alternatives.

Consider the statistic

$$Z = \frac{\bar{Y} - \bar{W} - \mu_1 + \mu_2}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}. \tag{12.9}$$

Then Z has a standard normal distribution and tests and confidence intervals may be based on that. For example, to test $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ and appropriate test is to reject H_0 if

$$|\bar{Y} - \bar{W}| > z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

where α is the desired size of the test.

12.3.3 Comparison of two normal means, variances common but unknown

Consider now the same situation as in the previous example, but suppose that σ_1^2 and σ_2^2 are unknown, but we do assume they are equal to a common value σ^2. We can estimate σ^2 by the combined sample variance

$$s^2 = \frac{\sum(Y_i - \bar{Y})^2 + \sum(W_i - \bar{W})^2}{m+n-2}$$

which has the distributional property

$$\frac{(m+n-2)s^2}{\sigma^2} \sim \chi^2_{m+n-2};$$

moreover, s^2 is independent of $\bar{Y}$ and $\bar{W}$. It follows that we may define

$$t = \sqrt{\frac{mn}{m+n}} \frac{\bar{Y} - \bar{W} - \mu_1 + \mu_2}{s}. \tag{12.10}$$

(compare equation (12.9)), and this quantity has a t_{m+n-2} distribution. Tests and confidence intervals for $\mu_1 - \mu_2$ may then be based on the statistic t defined in (12.10).

12.3.4 Comparison of two normal means, variances completely unknown

What happens if, in the context of the previous example, if σ_1^2 and σ_2^2 are not assumed equal? This is the famous *Behrens-Fisher* problem, named after W.-U. Behrens who first wrote about the problem in 1929, and R.A. Fisher who subsequently wrote about it at great length. The surprising fact is that this problem is vastly more complicated than the other ones we have been considering, and indeed does not have any solution of the same type as the others that we have developed. The problem can be solved if $r = \sigma_1^2/\sigma_2^2$ is known, and indeed the preceding subsection explains what to do if $r = 1$; the case where r is some other known value is only a little more complicated. Since it is also possible to construct tests and confidence intervals for r (Section 12.3.6 below), an *ad hoc* solution is to estimate r (or test whether $r = 1$) and then proceed as if r were known. However, this does not satisfy the exact probability requirements of a test or confidence interval.

An alternative procedure which can be applied when $m = n$ is that based on paired comparisons. Consider the differences $Y_1 - W_1$, $Y_2 - W_2$,..., $Y_n - W_n$; these are independent $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ so that the method of Section 12.3.2 may be used to form confidence intervals or hypothesis tests about $\mu_1 - \mu_2$.

Some variant of this idea is commonly applied in clinical trials. Suppose the random variables Y and W represent responses to two courses of treatment for a disease. It is possible to take pairs of patients, as closely as possible matched in terms of age, sex and disease condition, and then randomly assign one patient to receive one treatment and the other patient to receive the other. The resulting samples of patients receiving the two treatments will not be homogeneous, and an analysis involving paired comparisons is often appropriate.

However, this is a somewhat different situation from the one with which we started this section, which assumes that each of the two samples represents an independent sample of identically distributed observations. In this case the grouping of observations to form a paired comparison study will be totally arbitrary, and as a result information may be lost in the analysis. In more technical terms, a paired comparison analysis fails to satisfy the intuitive property that it should be *invariant* under permutations of the observations within each sample. However, a famous result due to Scheffé showed that this difficulty is inherent to the Behrens-Fisher problem: that there does not exist a procedure which is invariant under permutations of the observations within each sample, and which satisfies the exact probability requirement of a hypothesis test or confidence interval. The alternative proposed by Behrens and Fisher is known as fiducial analysis, but this lies outside the scope of the present discussion.

12.3.5 Estimation of a population variance

Suppose now we again have a single sample, $Y_1, ..., Y_n$ from $N(\mu, \sigma^2)$, and we are interested in estimating σ^2. The appropriate sampling statistic is the sample variance s^2, and Proposition 12.3 gives its distribution. For example, suppose we are interested in a $100(1-\alpha)\%$ confidence interval for σ^2. Defining $X = (n-1)s^2/\sigma^2$ we may write

$$\begin{aligned} 1-\alpha &= \Pr\{\chi^2_{n-1;\alpha/2} \le X \le \chi^2_{n-1;1-\alpha/2}\} \\ &= \Pr\left\{\chi^2_{n-1;\alpha/2}\cdot\frac{\sigma^2}{n-1} \le s^2 \le \chi^2_{n-1;1-\alpha/2}\cdot\frac{\sigma^2}{n-1}\right\} \\ &= \Pr\left\{\frac{(n-1)\cdot s^2}{\chi^2_{n-1;1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)\cdot s^2}{\chi^2_{n-1;\alpha/2}}\right\} \end{aligned}$$

so that

$$\left[\frac{(n-1)\cdot s^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)\cdot s^2}{\chi^2_{n-1;\alpha/2}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

Note, however, that there is one contrast between this calculation and the earlier ones about means. The standard normal and t distributions are both symmetric about their mean at 0, so it is natural to define a confidence interval in such a way that the error probability α is equally divided between the two tails, i.e. so that there is probability $\alpha/2$ that the true value lies to the left of the quoted confidence interval and probability $\alpha/2$ that the true value lies to the right. The χ^2 distribution is not symmetric, so there is no particular reason to follow this convention here. Indeed, it would be possible to construct slightly shorter confidence intervals by abandoning this requirement. However, the equal-tailed confidence intervals are the most natural and the easiest to construct, so it is usual to stick with them in practical applications.

As an example of these calculations, suppose $n = 10$ and we are interested in a 95% confidence interval. We have $\chi^2_{9;0.025} = 2.70$ and $\chi^2_{9;0.975} = 19.02$; moreover $9/2.70 = 3.33$ and $9/19.02 = 0.473$ so the 95% confidence interval runs from $0.473s^2$ to $3.33s^2$. For a 99% confidence interval, we have $9/\chi^2_{9;0.005} = 9/1.73 = 5.20$ and $9/\chi^2_{9;0.995} = 9/23.59 = 0.382$ so the confidence interval runs from $0.382s^2$ to $5.20s^2$. The considerable width of these confidence intervals is to some extent in contrast with the comparatively modest increase in the length of the confidence interval for a sample mean which is needed to allow for the estimation of σ^2 (recall the discussion at the end of Section 12.3.1).

12.3.6 Ratio of two normal variances

Now consider the same situation as in Sections 12.3.2–12.3.4, i.e. $Y_1, ..., Y_m$ and $W_1, ..., W_n$ are two independent samples from distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, and suppose our interest is in the ratio σ_1^2/σ_2^2. Calculate the sample variance s_1^2 and s_2^2 and define U by (12.3); then tests and confidence intervals may be based on the $F_{m-1,n-1}$ distribution of U.

As an example, suppose we wish to test $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$. Under the null hypothesis, U is just σ_1^2/σ_2^2 so the test is to reject H_0 if

$$s_1^2/s_2^2 < F_{m-1,n-1;\alpha/2} \quad \text{or} \quad s_1^2/s_2^2 > F_{m-1,n-1;1-\alpha/2}.$$

For example, suppose $m = 10$ and $n = 15$ and we again fix $\alpha = 0.05$. We find $F_{9,14;0.975} = 3.21$ and $F_{9,14;0.005} = 0.263$[1] and then we deduce that we should reject H_0 if s_1^2/s_2^2 is either less than 0.263 or greater than 3.21. Once again, it often seems surprising that such comparatively large or small ratios of s_1^2/s_2^2 should be considered consistent with the null hypothesis, but this again reflects the considerable uncertainty in estimating variances from such comparatively small samples.

[1]In S-PLUS or R, these percentage points may be obtained by typing `qf(0.975,9,14)` or `qf(0.025,9,14)` respectively. If using statistical tables, it may be necessary to look up $F_{14,9;.975} = 3.80$ and then deduce $F_{9,14;0.005} = 1/F_{14,9;.975} = 0.263$. Note also that some interpolation in the tables may be necessary to achieve these results.

12.4 Joint and conditional densities, and the multivariate normal distribution

12.4.1 Densities of random vectors

Consider the case of a p-dimensional random vector $Y = (Y_1, ..., Y_p)^T$. The density of Y at $y = (y_1, ..., y_p)^T$, denoted $f_Y(y)$, exists if the limit

$$f_Y(y) = \lim_{h_1 \downarrow 0, ..., h_p \downarrow 0} \frac{\Pr\{y_1 < Y_1 \leq y_1 + h_1, ..., y_p < Y_p \leq y_p + h_p\}}{h_1 ... h_p}$$

exists. Usually we consider the distribution of a random vector to be continuous if $f_Y(y)$ exists for every $y \in \mathcal{R}^p$, though it may be 0 for some y.

Suppose $Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix}$ where $Y^{(1)}$ consists of the first q components of Y and $Y^{(2)}$ consists of the last $p - q$. The marginal density of $Y^{(1)}$ is obtained by integrating out the components of $Y^{(2)}$,

$$f_{Y^{(1)}}(y^{(1)}) = \int f_Y \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} dy^{(2)} \qquad (12.11)$$

where the integral in (12.11) is typically over the whole of $\mathcal{R}^{p-q}$.

The conditional density of $Y^{(2)}$ given $Y^{(1)} = y^{(1)}$, denoted $f_{\{Y^{(2)}|Y^{(1)}=y^{(1)}\}}(y^{(2)})$, is defined by the formula

$$f_Y \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} = f_{Y^{(1)}}(y^{(1)}) f_{\{Y^{(2)}|Y^{(1)}=y^{(1)}\}}(y^{(2)}). \qquad (12.12)$$

Suppose Y is a p-dimensional random vector with density $f_Y(y)$, and let $Y = h(Z)$ for some differentiable one-to-one function h. The density of Z, denoted $f_Z(z)$, is given by

$$f_Z(z) = f_Y(h(z))\ |J|, \qquad (12.13)$$

where $|J|$ denotes the determinant of the matrix J, and J is the *Jacobian matrix* whose (i, j) entry is

$$J(i, j) = \frac{\partial y_i}{\partial z_j}.$$

In particular, for a linear transformation $Y = BZ$ for some nonsingular $p \times p$ matrix B, $J = B$ and so

$$f_Z(z) = f_Y(Bz)\ |B|. \qquad (12.14)$$

12.4.2 Means and Covariance Matrices

Suppose Y is a p-dimensional random vector with density $f_Y(y)$. The mean of Y, denoted $\mathrm{E}(Y)$ or μ_Y, is defined by

$$\mathrm{E}(Y) = \mu_Y = \int y f_Y(y) dy, \tag{12.15}$$

where the integral may without loss of generality be taken to be over the whole of $\mathcal{R}^p$ since any parts where Y is undefined may be taken to have $f_Y(y) = 0$.

For a discrete random variable which takes a countable set of values $\{y^{(i)}\}$ with probability mass function $p_Y(y^{(i)})$, the corresponding formula is

$$\mathrm{E}(Y) = \mu_Y = \sum y^{(i)} p_Y(y^{(i)}). \tag{12.16}$$

The covariance matrix of Y is defined by

$$\Sigma_Y = \mathrm{E}\{(Y - \mu_Y)(Y - \mu_Y)^T\}. \tag{12.17}$$

If $Z = AY + b$ is some linear transformation of Y, then

$$\begin{aligned} \mu_Z &= A\mu_Y + b, \\ \Sigma_Z &= A\Sigma_Y A^T. \end{aligned}$$

12.4.3 The multivariate normal distribution

In this subsection we state and prove a few of the elementary properties of the multivariate normal distribution with nonsingular covariance matrix. No attempt is made to be comprehensive; the objective is to provide necessary background for the (relatively few) places that this distribution is used in the text.

Suppose Y is a p-dimensional random vector with mean μ_Y and covariance matrix Σ_Y, and suppose Σ_Y is nonsingular. Y is said to have a *multivariate normal distribution* if it has the density

$$f_Y(y) = (2\pi)^{-p/2} |\Sigma_Y|^{-1/2} \exp\left\{ -\frac{1}{2}(y - \mu_Y)^T \Sigma_Y^{-1} (y - \mu_Y) \right\}. \tag{12.18}$$

Here are a few properties of the multivariate normal distribution.

Proposition 12.5. If Σ_Y is a diagonal matrix with diagonal entries $\sigma_1^2 > 0, ..., \sigma_p^2 > 0$, and if $\mu_Y = (\mu_1, ..., \mu_p)^T$, then the statement that Y has a multivariate normal distribution with mean μ_Y and covariance matrix Σ_Y is equivalent to the statement that $Y_1, ..., Y_p$ are independent random variables with $Y_i \sim N(\mu_i, \sigma_i^2)$.

Proof. If Σ_Y is diagonal then $|\Sigma_Y| = \prod_{i=1}^{p} \sigma_i^2$, and Σ^{-1} is also diagonal with diagonal entries $\sigma_1^{-2}, ..., \sigma_p^{-2}$. Therefore, (12.18) is equivalent to the density

$$f_Y(y) = \prod_{i=1}^{p} \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma_i}\right)^2 \right\} \right]. \tag{12.19}$$

But (12.19) is the product of $N(\mu_i, \sigma_i^2)$ densities, and therefore establishes that $Y_1, ..., Y_p$ are independent and normally distributed. The reverse argument is the same: if we are given that $Y_1, ..., Y_p$ are independent normal, then (12.19) is the joint density, but this is the same as the joint density (12.18) for the multivariate normal.

Proposition 12.6. If Y is multivariate normal with mean μ_Y and covariance matrix Σ_Y, and if $Z = AY + b$ with A a $p \times p$ *nonsingular* matrix and $b \in \mathcal{R}^p$, then Z is multivariate normal with mean $\mu_Z = A\mu_Y + b$ and covariance matrix $\Sigma_Z = A\Sigma_Y A^T$.

Proof. Write $Y = A^{-1}(Z - b)$. This is a one-to-one differentiable transformation with Jacobian matrix $J = A^{-1}$. By the transformation rule (12.13), the density of Z is

$$\begin{aligned}
f_Z(z) &= f_Y(A^{-1}(z-b))\ |A|^{-1} \\
&= (2\pi)^{-p/2}|\Sigma_Y|^{-1/2}|A|^{-1}\cdot \\
&\quad \cdot \exp\left[-\frac{1}{2}\{A^{-1}(z-b) - \mu_Y\}^T \Sigma_Y^{-1} \{A^{-1}(z-b) - \mu_Y\} \right] \\
&= (2\pi)^{-p/2}|\Sigma_Z|^{-1/2}\cdot \\
&\quad \cdot \exp\left\{ -\frac{1}{2}(z - b - A\mu_Y)^T (A^T)^{-1}\Sigma_Y^{-1}A^{-1}(z - b - A\mu_Y) \right\} \\
&= (2\pi)^{-p/2}|\Sigma_Z|^{-1/2} \exp\left\{ -\frac{1}{2}(z-\mu_Z)^T \Sigma_Z^{-1}(z-\mu_Z) \right\}.
\end{aligned}$$

This is of the form required for the result.

Remark. Proposition 12.6 is actually true in much greater generality than here stated — in particular, it is true without the assumption that A be nonsingular (but in that case needs to be interpreted differently, since Z does not have a density) and also in the case when A is a $q \times p$ matrix with q not necessarily equal to p. We have stated it in the simpler form here because this is all that is needed for the results in the text.

We note one consequence of Proposition 12.6, which is critical to the proofs of section 3.9:

Proposition 12.7. Suppose $Y = (Y_1, ..., Y_p)^T$ where $Y_1, ..., Y_p$ are independent $N(0, \sigma^2)$. Suppose $Z = (Z_1, ..., Z_p)^T = QY$, where Q is orthogonal, i.e. $QQ^T = Q^T Q = I$. Then $Z_1, ..., Z_p$ are also independent $N(0, \sigma^2)$.

Proof. $\mu_Z = Q\mu_Y = 0$, $\Sigma_Z = Q\Sigma_Y Q^T = \sigma^2 QQ^T = \sigma^2 I$. By Proposition 12.6, Z is multivariate normal with mean 0, covariance matrix $\sigma^2 I$. By Proposition 12.5, this means that $Z_1, ..., Z_p$ are independent $N(0, \sigma^2)$. Hence the result is established.