CHAPTER 15. CONFIDENCE REGIONS

In hypothesis testing, we begin with a fixed model and then use an observation to confirm or reject that model. It often happens, however, that we have no single model in mind to begin with, and that we would like to use the observation to help find a model that is likely to be close to the true one, or to help find a set of models that is likely to include the true model. In Chapter 12, we saw how estimation methods can be used to help indicate a single model. The present chapter describes an important and commonly used way of indicating a set of models. It is called the method of confidence regions. It forms (along with hypothesis testing and estimation) a third major area of classical mathematical statistics. The method of confidence regions is simple and easy to understand, and is closely related, in its concepts, to hypothesis testing.

In Chapter 14, we took a fixed model and asked which observations would give us, under that model, a DLS > α. The set of such observations was called the acceptance region determined by the given model at critical level α. If our actual observation fell in the acceptance region, then we accepted the hypothesis that the given model was correct. If not (that is, if the actual observation fell in the critical region), then we rejected that hypothesis. Note that the acceptance region was a set of possible observations.

Now, in the present chapter, we take a fixed observation and we ask what models would give us, for that observation, a DLS > α. The set of such models is called the confidence region determined by the given observation at critical level α. Note that the confidence region is a set of possible models.

Example. Consider a binomial experiment of 100 trials. We carry out the experiment and get $X = 45$ successes as our observation. Let $\alpha = 0.05$. What is the confidence region determined by this observation? The confidence region will be an interval of values of p. The endpoints of this interval will be those values of p for which the observation $x = 45$ has DLS $= 0.05$. We find these endpoints as follows. (We use normal approximation, but otherwise our calculation is exact).

For a given value of p, the DLS of $x = 45$ is obtained by normal approximation as

$$\text{DLS} = 1 - 2A(z), \qquad \text{where}$$

$$z = \frac{|45-100p| - \frac{1}{2}}{\sqrt{100p(1-p)}} .$$

For this DLS to be 0.05, we must have (as we know from normal tables) $z = 1.96$. Hence we must solve the equation

$$\frac{|45-100p| - \frac{1}{2}}{\sqrt{100p(1-p)}} = 1.96 .$$

We consider two cases. In the case $p < 0.45$, we have $45-100p > 0$. Hence we have

$$45-100p-\tfrac{1}{2} = 1.96\sqrt{100p(1-p)} .$$

Squaring, we obtain the quadratic equation

$$10.38p^2-9.28p+1.98 = 0$$

Solving, we find, as the solution below 0.45, $p = 0.351$. In the case $p \geq 0.45$, we have $100p-45 \geq 0$. Hence we have

$$100p-45-\tfrac{1}{2} = 1.96\sqrt{100p(1-p)}.$$

Squaring, we obtain the quadratic equation

$$10.38p^2-9.48p+2.07 = 0.$$

Solving, we get, as the solution above 0.45, $p = 0.552$. Thus, our desired confidence region is

$$0.351 < p < 0.552 .$$

Relation between acceptance region and confidence region. Fix a critical level α. Note that an observation ω lies in the acceptance region determined by a model μ if the DLS of ω under μ is $> \alpha$. Note also that a model μ lies in the confidence region determined by an observation ω if the DLS of ω under μ is $> \alpha$. Thus we immediately see that for fixed critical level α, the acceptance region for μ contains ω if and only if the confidence region for ω contains μ.

For the example of 100 Bernoulli trials, we saw in Chapter 13 that the acceptance region for the model $p = \frac{1}{3}$ is the interval $24 \leq X \leq 43$. Above, we have seen that the confidence region for the observation $X = 45$ is the interval $0.351 < p < 0.552$. This is in accord with the above principle relating acceptance regions and confidence regions; for $X = 45$ is not in the acceptance region for $p = \frac{1}{3}$, and $p = \frac{1}{3}$ is not in the confidence region for $X = 45$.

Terminology. In the example above, the models were values of p. The confidence region was an interval of values of p. In such a parametric case, the confidence region is called a confidence interval. The endpoints of the interval are called confidence limits. For a confidence region, the quantity $\gamma = 1-\alpha$ is called the confidence level.

In the example above, the confidence level is $\gamma = 1-0.05 = 0.95$. Usually, confidence levels are taken at 0.95 or 0.99. We then speak, for example, of a 0.95 confidence region. (In the same way, observations in an acceptance region may lie in an interval. We may then speak of such an interval as an acceptance interval and of numbers giving endpoints for it as acceptance limits.)

We emphasize again that an acceptance region is a set of possible observations, while a confidence region is a set of possible models.

Meaning of the confidence level γ. Let μ be the true (but unknown) model for an experiment. Let $\alpha = 0.05$. Then $\gamma = 1-\alpha = 0.95$. Repeat the experiment many times. Then approximately $\alpha (= 0.05 = 5\%)$ of all the observations obtained will fall outside the acceptance region for μ. (This was the meaning for the critical level α.) Hence approximately $\gamma (= 0.95 = 95\%)$ of all observations will lie in the acceptance region. Take a confidence region for each observation. By the above principle, relating acceptance regions and confidence regions, approximately $\gamma (= 0.95 = 95\%)$ of these observations will have μ in their

confidence regions. Thus we have the following. If we make many independent repetitions of the process of getting an observation and calculating a confidence region, then in approximately $\gamma (= 0.95)$ of these cases, the resulting confidence region will include the true model. Similarly for other values of α and γ (such as $\alpha = 0.01$ and $\gamma = 0.99$). Note that as α becomes small, both the acceptance region and the confidence region increase in size.

It is sometimes tempting to say, "the probability is 0.95 that the true model μ lies in this confidence interval." This is incorrect, since the true model μ is fixed and cannot vary as the experiment is repeated. To be correct, we must say that "the probability is 0.95 that a confidence interval obtained in this way will include the true model."

Tables. Confidence intervals for binomial experiments have been tabulated. These tables are often given in graphical form by figures such as the following.

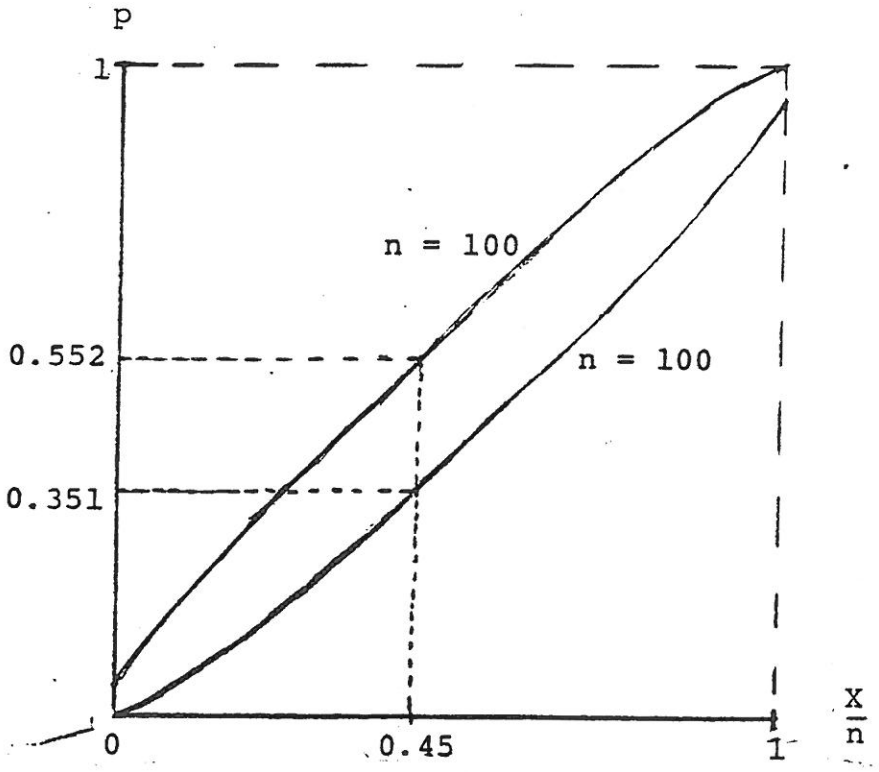

Figure 15.1
(for confidence level = .95)

Here observations are given on the horizontal axis (as values of $\frac{X}{n}$) and models are given on the vertical axis (as values of p). The two curves give the upper and lower confidence limits for $n = 100$. Other curves for other values of n could also be given in the same figure.

Example. Consider, as a second example, a Poisson experiment where a single trial is made. A model is given by a value for the parameter m in the Poisson formula. An observation is given by X, the number observed in the single trial of the experiment. Since the value of m is also, as we noted in Chapter 7, the

average value that we expect to observe in many repetitions of a Poisson experiment, we take m to be our theoretically expected result, and then, to get the DLS, we use the metric $|x-m|$.

How would we calculate the DLS for the observation $X = 0$ under the model $m = 2$? Here $|x-m| = 2$. The observations that lie at least as far as this from m are the values $X = 0$ and $X \geq 4$. From tables, we get the DLS $= P(X = 0 \text{ or } X \geq 4) = 0.278$.

Now let us take $\alpha = 0.05$. What is the acceptance region for $m = 2$? From tables, we get (as above) that the DLS for $x = 5$ is 0.053, and that the DLS for $x = 6$ is 0.017. Hence the acceptance region for $m = 2$ is the set of integers x, $0 \leq x \leq 5$.

What is a confidence region, again with $\alpha = 0.05$ (and hence confidence level $\gamma = 0.95$), for the observation $X = 0$? From tables we get that with $m = 5$, the DLS of $x = 0$ is 0.039, while with $m = 4.5$, the DLS of $X = 0$ is 0.051. Interpolating, we estimate that for $m = 4.54$, the DLS of $X = 0$ is 0.050. Hence we have as our confidence interval

$$0 \leq m < 4.54.$$

In the same way we can verify, for example, that the acceptance region for $m = 5$ is the set of integers x with $1 \leq x \leq 9$, and that the confidence region for $X = 5$ is $1.97 < m < 12.3$.

Approximation in the binomial case. In the case of a binomial experiment, where X is the observed number of successes, it is sometimes convenient to express the observation in the form $\frac{X}{n}$, the observed relative frequency of successes. The observation then lies in the interval $0 \leq \frac{X}{n} \leq 1$. Each model p is also given by a number in the interval $0 \leq p \leq 1$. The fact that observations and

models can both be pictured as lying in the interval from 0 to 1 can be a source of confusion when we consider acceptance regions and confidence regions for binomial experiments. An added source of confusion is that for any given value r between 0 and 1 and for any $\alpha > 0$, acceptance limits (for values of $\frac{X}{n}$) for a binomial experiment with the model $p = r$ and the confidence limits from the observation $\frac{X}{n} = r$ are usually very nearly the same. This is a special feature of the binomial distribution. (Note that in the Poisson example above, the confidence limits from $x = 5$ and acceptance limits with $m = 5$ were quite different.)

Thus, in our first example above, we found confidence limits for the model p from the observation $\frac{X}{n} = 0.45$ to be 0.351 and 0.552. If we now get acceptance limits for the observation $\frac{X}{n}$ from the model $p = 0.45$, we obtain, by normal approximation, 0.347 and 0.553. Since $n = 100$, the acceptance region consists of values of $\frac{X}{n}$ from 0.35 to 0.55.) Note that the limits are not the same, although they are very close. The mathematical procedures in the two cases are distinct. In both cases (for $\alpha = 0.05$, for example) we start with the equation

$$\frac{|x-np|-\frac{1}{2}}{\sqrt{np(1-p)}} = 1.96,$$

which can be rewritten as

$$\frac{\left|\frac{X}{n}-p\right|-\frac{1}{2n}}{\sqrt{\frac{p(1-p)}{n}}} = 1.96.$$

For acceptance limits, we fix p and solve for X (or $\frac{X}{n}$). This amounts to solving certain linear equations. For confidence limits, we fix X and solve for p. This amounts (as we saw above) to solving certain quadratic equations.

It is often convenient, however, to use a simpler approximating formula in both cases. We take the last formula above and drop $\frac{1}{2n}$ (which becomes small as n becomes large). This gives us

$$\left|\frac{X}{n} - p\right| = 1.96 \sqrt{\frac{p(1-p)}{n}}$$

Thus, for acceptance limits, we have

$$\frac{X}{n} = p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$$

For confidence limits, we take the further approximating step of using the observed value $\frac{X}{n}$ to approximate p under the square root sign. Letting $\hat{p} = \frac{X}{n}$, we have, for confidence limits,

$$p = \frac{X}{n} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad .$$

Note the similarity of the last two formulas. For $p = 0.45$ and for $\frac{X}{n} = 0.45$, they give identical limits: 0.352 and 0.548. (If we ask for acceptance limits to be observable values, we get 0.36 and 0.54. The change from the previously found limits 0.35 and 0.55 is caused by the omission of the correction term for bar width.)

This coincidence, which is special for the binomial case, can cause confusion. It is important to keep in mind that we are working, conceptually, with two separate sets: the universe of possible models and the set of possible observations. They will be clearly different, as sets, in almost all situations other than binomial experiments.

Example. Out of 1000 randomly selected cars, 100 are observed to be white. Find approximate confidence limits for the percentage of white cars in the population from which the sample is taken.

Here $\frac{X}{n} = 0.1$. Using the simpler formula and letting $\hat{p} = 0.1$, we have, as confidence limits

$$p = 0.1 \pm 1.96 \sqrt{\frac{(0.1)(0.9)}{1000}}$$

$$= 0.1 \pm 0.019 .$$

Thus the confidence limits for the percentage of white cars are 8.1% and 11.9%.

Confidence regions and hypothesis tests. If we are given a model as null hypothesis, an observation, and a critical level α, we can use the observation to carry out a hypothesis test on the model. If the model is rejected, the hypothesis test procedure, by itself, does not tell us how far away from the null hypothesis the correct model may lie. On the other hand, if we also use the observation to construct a confidence region, then this confidence region does

give an indication of where, in the universe of models, the true model may lie. Construction of a confidence region thus gives more information than a simple hypothesis test. This is not surprising, since a hypothesis test only considers the relation of the observation to the null hypothesis, while the confidence region takes into account the relation of the observation to all models in the universe. Although they are more useful, confidence regions are, in general, more difficult to construct than the acceptance region for a hypothesis test. Design of a hypothesis test requires only that we have a metric for the null hypothesis, while construction of a confidence region requires that we have a form of metric that can be applied to each model in the universe. In some cases, there may be no evident way to choose a metric for models other than the null hypothesis. Here a hypothesis test may be the only possible approach. This is the case, for example, with contingency tables of type α, and with other uses of the CS-metric to test goodness-of-fit.

The case of composite models also presents difficulties for making confidence regions. In special cases, as we shall see in Chapter 16, it may be possible to think of the entire universe as a set of non-overlapping composite models and to find a form of metric that is well-defined for each of these composite models. Construction of a confidence region in this case then takes the form of the construction of a certain collection of these composite models.

Constructing confidence regions. A confidence region can be constructed from an observation as soon as we decide on (i) a universe of models, (ii) a form of metric that can be applied to each model in the universe, and (iii) a confidence level. In many cases, the choices of universe and metric will be natural and evident. Construction of the region then follows directly as soon as we choose a confidence level. We look at a further example.

Example . Each year, dog licenses in a certain town are numbered consecutively beginning with 1. A single dog is chosen at random and observed to have license number 42. Find 95% confidence limits for the total number N of licenses that have been issued in that year. This problem can be solved as follows. The possible values of N are $\{42,43,\ldots\}$. Each value of N gives a model in which the possible observations $\{1,2,\ldots,N\}$ are equiprobable. Consider an observation $X = x_o$ and the model given by N_o. Then $s(X) = N_o - X$ is a useful metric to measure the distance of X from giving strongest confirmation of the model. Under the model N_o, the DLS of $X = x_o$ will then be the probability that $X \leq x_o$ for another observation X, and this probability will be $\frac{x_o}{N_o}$. Hence, if $X = x_o = 42$ and $\gamma = 0.95$, we seek those N_o such that the DLS of x_o is > 0.05. We thus have

$$\frac{42}{N_o} > 0.05.$$

Solving this inequality, we get $N_o < 840$. Therefore the desired confidence limits are 42 and 839.

APPENDIX TO CHAPTER 15.

To illustrate the logical process of constructing a confidence region, we give a further and more subtle example.

Example. Consider the one-sided universe given in the binomial example in Chapter 14 for a one-sided test. In that example (concerning sterilization of bandages), the universe consisted of values of p such that $1/3 \leq p \leq 1$. To get confidence intervals, the metric given there for the model $p = 1/3$ must be extended to other models in the universe. One way of doing this is to let

$$s(x;p) = \begin{cases} |x-np| & \text{when} \quad \frac{n}{3} \leq x, \\ \\ |\frac{n}{3} - np| & \text{when} \quad x < \frac{n}{3}. \end{cases}$$

(In the example in Chapter 14, we have $n = 100$.)

This form of metric puts all observations to the left of $\frac{n}{3}$ at the same distance from np as $\frac{n}{3}$ itself. It follows, as can be checked, that for observations above $\frac{n}{3}$, the upper confidence limit is the same as in the two-sided case, and that for observations below $\frac{n}{3}$, the upper confidence limit has the same value as it would for an observation $\approx \frac{n}{3}$. The lower confidence limit is never less than $\frac{n}{3}$, and, for sufficiently large observations, it is larger than the lower confidence limit for the same observation in the two-sided case.

Take, for example, the observation $X = 45$ in a binomial experiment with 100 trials. In the two-sided case, with the universe $0 \leq p \leq 1$, the 0.95 confidence limits are 0.351 and 0.552 as we saw earlier in this section. In the one-sided case, with the universe $\frac{1}{3} \leq p \leq 1$, the upper limit is 0.552 and the lower limit is 0.366. (The reader may check that the DLS of $x = 45$ for the model $p = 0.366$ is 0.05 under the one-sided metric $s(x;p)$ defined above.) In this example, other choices of metric are also possible.

In each case, the 0.95 confidence intervals constructed have the desired property that, when construction of an interval is repeated for many successive observations, the constructed interval includes the true model about 95% of the time. The reader should note, in the example, that in going from the two-sided to the one-sided case, the shorter intervals for large observations exactly compensate for the larger upper confidence limit for small observations in such a way as to preserve the 95% confidence for all models $p \geq 1/3$. In that example, the individual confidence intervals may be viewed as altered (from the two-sided case) by the additional information that p must be $\geq 1/3$.

Remark. The confidence region that we construct from a given observation under a given form of metric will depend upon the metric. What makes one metric preferable to another for the purposes of constructing confidence regions? We consider this further in our general discussion of metrics in Chapter 20.

EXERCISES FOR CHAPTER 15

15-1. In a sample of 20 students drawn from a large population, 16 recalled recently-learned material better immediately after sleeping 8 hours than after 8 hours awake. Set 80% confidence limits on the population proportion p who would have performed better after sleeping 8 hours, had all been tested.

15-2. A random sample of 50 families from Cambridge, Massachusetts, showed 10 families with incomes over \$10,000 during 1969. Set a 90% confidence interval on p, the percent of families with incomes over \$10,000.

15-3. Among 211 cardiac invalids in a series having heart operations, the operative mortality was 18%. Set 95% confidence limits on the operative mortality for this operation on this kind of patient.

15-4. Consider a binomial experiment with $n = 100$. Consider the model $p = 0.6$ and the observation $X = 70$.

a. Find the DLS of this observation under this model.

b. Find the acceptance region determined by the model $p = 0.6$ when the critical level $\alpha = 0.01$.

c. Find the confidence interval determined by the observation $X = 70$ when the confidence level is 0.99.

15-5. Consider a Poisson experiment where a single observation is made. Consider the model $m = 4$ and the observation $X = 8$.

a. Find the DLS of this observation under this model. (See Exercise -12.)

b. Find the acceptance region determined by the model when $\alpha = 0.05$.

c. Find the confidence interval determined by the observation when confidence level is 0.95.

15-6. Consider the following test of a vaccine for the common cold. One group of 100 people is given the vaccine. Another group of equal size is given a placebo. If, after a mild winter, the results were

	vaccine	placebo
Caught cold	10	20
Cold free	90	80
	100	100

Use the observation to test for association, with critical level $\alpha = 0.05$.

CHAPTER 16. NONPARAMETRIC METHODS.

In the previous two chapters, we studied the forms of statistical analysis known as hypothesis testing and constructing confidence regions. These forms of analysis can be applied to a variety of new situations, provided that we can do the following: (a) define the experiment that we will conduct and the nature of the observation that we will obtain ; (b) identify (for the case of a hypothesis test) the model (or composite model) that we wish to test; (c) identify (for the case of a confidence region) the universe of models that we will use; (d) decide on a form of metric to measure distance of a given observation from highest confirmation of a given model.

The choice of a metric (d) is, perhaps, the key step and determines both the ease and the usefulness of the resulting analysis. Many metrics have been studied by statisticians, and, in a given situation, there may be several different and familiar metrics that can be applied. It is often possible for the beginner to discover an appropriate, useful, and natural metric by thinking carefully about the given circumstances. Several examples of this will follow. In each case, we shall be led to what is, in fact, a familiar and important metric in mathematical statistics.

When we look at a possible candidate metric (for a given problem) and seek to explore its usefulness, it is helpful to ask the following questions: (a) How do we calculate DLS values for the metric? Is the calculation simple? Are there approximations that can be used? Are there other useful short-cuts in

the computation? (b) Is it obvious from the method of calculating the DLS that there are composite models for which the metric is well-defined? (This often happens as we shall see.) (c) Does the metric measure distance in a way that seems appropriate to the practical circumstances of our problem and to the purposes of our work? (We approach this in a more precise way in Chapter 20.) (d) Does the metric lead us to powerful enough hypothesis tests? (To put it another way, does the metric make good use of the information contained in the observation?)

In this chapter we shall look at several examples of a search for a metric. In each case, the universe of models will be nonparametric. Hence, these examples will also serve as an introduction to the general subject of nonparametric methods.

The median of an observation. If we observe n independent values of a random variable and rank the observed values in order of size, and if we then take the middle value (if n is odd) or the average of the two middle values (if n is even), this result is called the median of the observation. This concept of median of a set of observed values should not be confused with the concept median of a random variable (described in Chapter 8). The two concepts are distinct: one is got from an observation; the other is got from a model (that is, from a distribution (density function)).

The median metric. Let a continuous random variable X have a given fixed distribution. Let m be the median of the random

variable X. We observe n independent values $X_1, X_2, \ldots, X_n$ of X. We call this set of values the observation Ω. We seek a metric for observations of this kind under the given model. One possibility that suggests itself is the following. Let m_Ω be the median of the observation Ω, and take $s_1(\Omega) = |m_\Omega - m|$ as the metric. If the assumed model is corrent and if n is large, we would (by the stability of relative frequencies) expect m_Ω to be close to m. Hence $s_1(\Omega)$ would seem a natural choice of metric. To calculate a DLS value for a given observation Ω requires, however, that we know the mathematical form of the density of X. For example, if the density is narrow (has small variance), we will get smaller DLS values for a given value of $S_1(\Omega)$ than if the density is more spread out (has large variance).

What if we do not know the exact form of the density but only know the value of the median m ? Can we find a useful metric where calculation of DLS values will only use the value of m and will not require further information about the density function? Consideration of specific examples will lead us to such a metric. Let X be a random variable with median m = 4.6. We take an observation of five values and get 2.4, 7.2, 3.0, 2.9, and 6.5. We then take a second observation of five values and get 5.5, 9.4, 6.1, 7.2, and 6.5. Which of these two observations seems closer to what we would expect from a distribution with median 4.6 ? Clearly the first observation does, since, in the first observation, roughly equal numbers of the observed values occur above and below the assumed median, while in the second observation, all the observed

values occur above the assumed median. With 4.6 as the assumed median, the second observation would appear to be less likely. We can make these ideas more precise by regarding each observed value as the result of a Bernoulli trial with <u>falls above 4.6</u> as success and <u>falls below 4.6</u> as failure. The probability p of success for each trial must be $p = P(4.6 < X) = P(m < X) = 1/2$ by the definition of median. Hence the first observation gives 2 successes in 5 Bernoulli trials (with $p = 1/2$), while the second observation gives the less likely event of 5 successes in 5 Bernoulli trials (with $p = 1/2$). Note that these results depend only on the median value 4.6 and do not use any further information about the distribution of X.

This leads us to define the following metric. A random variable X with median m is given. Then for each observation $\Omega = (X_1,\ldots,X_n)$ obtained as above, we let N_Ω be the number of observed values in Ω which fall above m. We define our metric to be

$$s^M(\Omega) = \left|N_\Omega - \frac{n}{2}\right| ,$$

and we call this metric the <u>median metric</u>. Values of the <u>DLS</u> can now be obtained by noting that N_Ω must have the binomial distribution for n trials with $p = 1/2$. (Indeed, if we think of N_Ω as the result of a binomial experiment, $s^M(\Omega)$ is the usual binomial metric.) Thus, in the specific examples just above, the <u>DLS</u> of the first observation is the DLS of 2 successes in 5 trials (and this <u>DLS</u> = 1), while the <u>DLS</u> of the second observation is the <u>DLS</u> of 5 successes in 5 trials (and this <u>DLS</u> = 2/32 = 1/16 = 0.06). Values of the <u>DLS</u> can

hence be obtained by direct calculation of the binomial distribution with $p = 1/2$, or by normal approximation to this binomial distribution.

Example. Assume that a random variable has median 4.0. 50 values are observed and 15 of them lie above 4.0. What is the DLS of this observation under the median metric? It is enough to find the DLS for 15 successes in a binomial experiment with 50 trials and $p = 1/2$. Using normal approximation, we have:

$$\text{DLS} = P(|N_\Omega - 25| \geq 10) = 1 - 2A(z),$$

where

$$z = \frac{10 - 1/2}{\sqrt{50/4}} = 2.69 .$$

Using a normal table, we get DLS = 0.007.

Composite models and the median metric. Since DLS calculations for the median metric depend only on the value of the median of the assumed distribution for X, we can view the set of all distributions with the same median as a composite model, and we see that the median metric is well defined for this composite model. Indeed, we can view the universe of all continuous distributions as formed of non-overlapping composite models - one composite model for each possible median value. Then we have that the median metric for each composite model is well defined with respect to that composite model. Thus we can use the median metric to conduct a hypothesis test as follows.

Example. A random variable X is given. We observe 10 independent values of X and get -5.6, -5.5, 7.4, 25.6,

-12.9, 12.5, -7.6, -11.5, -17.7, -0.6. Test the hypothesis that the median of X is -12.0, using significance level 0.05.

Solution. Call the given observation ω. 8 of the 10 observed values fall above -12.0. Hence $N_\omega = 8$. Hence the DLS will be

$$P(N_\Omega = 0,1,2,8,9, \text{ or } 10) = 2\left(\frac{1}{2^{10}} + \frac{10}{2^{10}} + \frac{45}{2^{10}}\right) = \frac{56}{2^9} = 0.11 ,$$

and we continue to accept the hypothesis.

Confidence intervals and the median metric. Because the median metric is defined for all models and observations, and because it is well defined on each composite model given by a value for the median, we can use it to form confidence regions, where each confidence region is given in the form of a set of composite models - that is to say, in the form of an interval of possible values of the median. Given an observation ω, we simply give, as our confidence interval, those values of m for which the DLS of ω is sufficiently high. Thus, taking the observation ω in the last example above, we get a 95% confidence interval for the median as follows. We seek those values of m for which the DLS of ω is > 0.05. When will the DLS be > 0.05 ? We saw above that $N_\omega = 8$ gives DLS = 0.11. We can also show that $N_\omega = 9$ would give DLS = 11/512 = 0.02. $N_\omega = 2$ gives the same DLS as $N_\omega = 8$, and $N_\omega = 1$ gives the same DLS as $N_\omega = 9$. Hence the confidence interval from ω will be formed of those values of m which give values of N_ω ranging

between 2 and 8. But it is evident that for N_ω to range between 2 and 8, values of m must range between the second and ninth largest of the observed values. In our example, these values are -12.9 and 12.5, hence the interval (-12.9,12.5) is the desired 95% confidence interval for m. (Note that if the number n of observed values is too small, we may not be able to get a finite confidence interval. When $n = 5$, for example, all values of m lie in the 95% confidence region, since the DLS values for $N_\omega = 0$ and for $N_\omega = 5$ are both 0.06.)

Remark. If we are using the median metric for a hypothesis test for a fixed value of m, how do we treat an individual observed value which happens exactly to equal m? Do we consider it as falling above m or below m? One rule (which we do not justify here) is as follows: if there is a single such value in ω, count it in the way that tends to confirm m. If there are several such values, count the first on the confirming side, the second on the disconfirming side, and continue to alternate. A count on the confirming side is a count which reduces the value of $S^M(\omega) = |N_\omega - \frac{n}{2}|$. If $N_\omega = \frac{n}{2}$, the choice is arbitrary. If there are many such values, it may be that the level of accuracy of the observations is not sufficiently high for the purposes of the test or that the assumption of continuous distributions is incorrect. (Remember that we are assuming a continuous random variable. According to such a model, the probability of observing any preassigned value of X must be extremely small since for any given interval $[a,b]$,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx.)$$

Remark. The median metric is most useful when the universe of models is the set of all continuous distributions. If a smaller universe is assumed, it may be possible to get other metrics which are better in that they give more powerful hypothesis tests and smaller confidence regions (for the same confidence level and the same observed data). In Chapter 17 we shall see that this is the case when we take the universe to be the set of all normal distributions. For that parametric universe, we will find a metric (Student's metric) which is also well defined with respect to (the composite model given by) each possible value of the median but which gives more powerful tests and smaller confidence regions than the median metric. (For the data in the example above, Student's metric gives [-11.04,7.86] as a 95% confidence interval in place of [-12.9,12.5].) DLS calculations for this metric (in this normal universe) are easy to carry out as we shall see. We cannot use Student's metric in the larger universe of all continuous distributions, since this metric will no longer be well defined on the desired composite models and, since, for non-normal models, the DLS calculation for Student's metric would be much more difficult.

The fact that reducing the size of the universe may give us better metrics does not come as a surprise. As we noted in Chapter 15 in connection with "one-sided" parametric universes, the knowledge (or assumption) that we are in a smaller universe serves as additional information, and we can expect this information to make our statistical analysis more discriminating.

The Wilcoxon-Mann-Whitney metric. Consider two continuous random variables X and Y with unknown distributions. A common form of statistical problem is the following: from observed data, decide whether X and Y have the same distribution. This problem often arises in biological science, for example, when X is a quantity (such as blood level of a certain substance) to be observed in an individual selected at random from a population that has been treated in some way, and Y is the same quantity to be observed from an untreated (control) population. We then ask whether the distribution for X (the treated group) is the same as the distribution for Y (the control group). (We considered the same problem in our study of contingency tables of type γ in Chapter 13. In that study, we classified the entire sample into two categories according to size of X and Y values, and we then used the CS-metric. In our present study, we shall introduce a new metric which leads to a more powerful and discriminating statistical analysis.

We observe m independent values of X $(x_1,\ldots,x_m)$ and n independent values of Y $(y_1,\ldots,y_n)$. Let ω be the observation $(x_1,\ldots,x_m;y_1,\ldots,y_n)$. A model for our experiment will be a pair of distributions, the first a distribution for X and the second distribution for Y. Let us now take the composite model M_o to consist of all models (pairs of continuous distributions) in which the two distributions are identical. Can we find a metric that is well-defined on M_o and will measure how far any observation ω is from giving strongest confirmation of M_o ?

We discover such a metric by asking ourselves: what sort of observation would tend to disconfirm the assumption of identical distributions? The simplest answer is: an observation in which the observed Y-values tend to fall above the observed X-values or else tend to fall below the observed X-values. How can we measure this tendency? One way is to look at the individual observed X and Y cases in all possible pairs (x_i, y_j) and see in how many cases $y_j - x_i$ is > 0 and in how many cases $y_j - x_i$ is < 0. (We assume for the moment that the X values are all distinct from the Y values, so that we never get $x_i = y_j$.) There are mn such pairs, and if X and Y have the same distribution, we would expect the number of positive differences to be roughly equal to the number of negative differences. Let U_ω be the number of positive differences among the mn pairs from the observation ω. Then we expect U_ω to be close to $\frac{mn}{2}$. This suggests that we take $|U_\Omega - \frac{mn}{2}|$ as our metric. This metric will be useful if we can show that resulting values of the DLS do not depend upon the specific form of the common distribution for X and Y (that is to say, if we can show that the metric is well defined on the composite model M_o of all identical pairs of continuous distributions.) We shall show this below. We call this metric the Wilcoxon-Mann-Whitney metric. (WMW metric). (Wilcoxon discovered the metric, Mann and Whitney simplified the calculations for the DLS, and the hypothesis test in which the metric is most commonly used is called the Mann-Whitney test.)

To see that the WMW metric is well defined on the composite model M_o of all identical pairs of continuous distributions, we need only note that if the entire set of $m+n$ observed values is ranked in order of size, and if the distribution for X is identical with the distribution for Y, then the various possible sets of positions of the Y values in this ordering are equally likely (since there is no reason for one set of positions to occur in preference to another.) For example, if $m = 2$ and $n = 2$, there are $\binom{4}{2} = 6$ possible sets of positions: XXYY, XYXY, YXXY, XYYX, YXYX, YYXX; and each has probability 1/6 of occurring. More generally, there are $\binom{m+n}{n}$ possible sets of positions for the Y values in the rank ordering of the $m+n$ observed values. Note that each set of positions (for possible Y values) gives a corresponding value of U. For example, with XYXY, $U = 3$ since $Y_j - X_i > 0$ for three of the $mn = 4$ pairs of X and Y values. We can now calculate the DLS of an observed U_ω by listing all sets of positions and then finding the number ν of sets of positions for which the value of $|U_\Omega - \frac{mn}{2}|$ is $\geq$ the observed value $|U_\Omega - \frac{mn}{2}|$. The DLS is then $\nu/\binom{m+n}{n}$.

Example. Two values (3.7,1.4) of the variable X and three values (1.6,4.2,3.9) of the variable Y are observed. If we assume identical distributions for X and Y, what is the DLS of this observation under the WMW metric? Here $m = 2$, $n = 3$, $U_\omega = 5$, and $|U_\omega - \frac{mn}{2}| = |5-3| = 2$. There are $\binom{5}{3} = 10$ possible sets of positions for the Y values. Tabulating these, with their corresponding values of U and $|U_\Omega - \frac{mn}{2}| = |U_\Omega - 3|$, we get:

Positions (Ω)	U_Ω	$\lvert U_\Omega - 3 \rvert$
XXYYY	6	3
XYXYY	5	2
YXXYY	4	1
XYYXY	4	1
YXYXY	3	0
YYXXY	2	1
XYYYX	3	0
YXYYX	2	1
YYXYX	1	2
YYYXX	0	3

4 of these sets of positions give $|U_\Omega - 3| \geq |U_\omega - 3| = 2$. Hence, the DLS of ω is $4/10 = 0.4$.

In applications, the values of m and n are usually somewhat larger than in the above example, and calculation of the exact DLS by the enumeration of all cases, as above, can be lengthy. Tables exist which give, for various pairs of values of m and n, and for various critical values α (such as $\alpha = 0.05$ and $\alpha = 0.01$), corresponding values for $|U_\omega - \frac{mn}{2}|$ at which the DLS goes below α. Such tables are not necessary, however, for two-decimal-place accuracy, because when $m = 1$ or $n = 1$ and when $m = 2$ or $n = 2$, the enumeration of possible cases is easy, and when $m \geq 3$, $n \geq 3$, and $m + n \geq 10$, the standard normal curve can be used to get two-decimal place accuracy in a way that we describe below. The remaining cases

($m \geq 3$, $n \geq 3$, $m + n \leq 9$) are not difficult to treat by enumeration.

Calculation of values of the metric $|U_\Omega - \frac{mn}{2}|$ for a given observation ω can be simplified as follows (as discovered by Mann and Whitney). The entire set of $m + n$ observed values is arranged in order of size. Each value is assigned a rank number (with the smallest value getting rank 1.) The sum of the rank numbers for the Y values is then calculated. We call the result the rank sum T_ω for Y. It is easy to show by a simple inductive proof that $T_\omega = U_\omega + \frac{n(n+1)}{2}$. It follows that $|U_\omega - \frac{mn}{2}| = |T_\omega - \frac{n(n+1)}{2} - \frac{mn}{2}|$. It is usually simpler to calculate T_ω first, and then obtain U_ω from T_ω. The proof of the identity $T_\omega = U_\omega + \frac{n(n+1)}{2}$ is as follows. Observe that when all Y values precede all X values, $U_\omega = 0$ and $T_\omega = \frac{n(n+1)}{2}$, so the identity holds. Any other ordering can be obtained from this ordering by successively interchanging pairs of adjacent X and Y values. But each such interchange either increases both T_ω and U_ω by 1 or decreases both T_ω and U_ω by 1. Hence the identity must continue to hold as the interchanges are made. Hence the identity holds for all orderings.

Normal approximation for the WMW metric. When m and n are sufficiently large (see above), the standard normal curve can be used to find approximate values of the DLS for the WMW metric as follows.

$$\text{DLS}(\omega) = P(|U_\Omega - \tfrac{mn}{2}| \geq |U_\omega - \tfrac{mn}{2}|) = 1 - 2A(\zeta),$$

where $$\zeta = \frac{|U_\omega - \frac{mn}{2}| - 1/2}{\sqrt{\frac{1}{12}mn(m+n+1)}}.$$

[Since the distribution for U_ω is symmetrical, values of the DLS for a one-sided version of the WMW metric can be found by taking the same ζ and finding DLS $= \frac{1}{2} - A(\zeta)$.] We shall derive the formula for ζ in Chapter 19. As with other metrics, much of the usefulness of the WMW metric comes from the fact that, for sufficiently large observations, DLS values can be conveniently approximated by the use of a known distribution (in the WMW case, the standard normal distribution) after a suitable change of scale.

The Mann-Whitney test. The hypothesis test which assumes the composite model M_o (consisting of all identical pairs of continuous distributions for the random variables X and Y) and which uses the WMW metric is called the Mann-Whitney test.

Example. Six college-bound students in a certain high school take a vocabulary review tutorial, and three do not. Scores of the group of six on a verbal aptitude test are: 585, 590, 609, 614, 622, and 625. Scores of the other three are: 576, 600, 606. Use the Mann-Whitney test at critical level 0.05 to decide whether the tutored students did significantly differently.

Solution. We let X = an untutored score and Y = a tutored score. Our null hypothesis is that X and Y have the same distribution. Our first decision is whether to apply the test as a one-sided test or as a two-sided test. If the

six students had been chosen for tutoring in a random way, then we could argue that the two groups were the same (had the same distribution as random variables) except for the effects of tutoring. We could then argue that tutoring could only act to improve scores and that we should apply a one-sided test. On the other hand, in the absence of random selection, it is possible that a process of self-selection for tutoring might lead to a tutored group that was, even after tutoring, inferior to the untutored group. This suggests that we should use a two-sided test, and we do so. The ordering of the entire sample observed is XYYXXYYYY. This gives the value $U_\omega = 14$. (Or we could calculate $T_\omega = 35$ and then get $U_\omega = 35 - \frac{6.7}{2} = 14$.) Hence the WMW metric has the value $|U_\omega - 9| = |14-9| = 5$. Enumerating cases, we get $\underline{DLS}(\omega) = P(U_\Omega \leq U_\omega) = \frac{2+2+4+6+8}{\binom{9}{3}} = \frac{22}{84} = 0.26$. Hence, we continue to accept the null hypothesis. (Normal approximation would give

$$\zeta = \frac{|14-9| - 1/2}{\sqrt{\frac{18.10}{12}}} = 1.16$$

and $\underline{DLS}(\omega_o) = 1 - 2A(\zeta) = 1 - 0.75 = 0.25$. With $m + n = 9$, the conditions given for two-decimal place accuracy of normal approximation are not quite satisfied.)

Remark. Other simple metrics also suggest themselves for the comparison of two random variables X and Y. One such metric would be $|m_\Omega^Y - m_\Omega^X|$ where m_Ω^X and m_Ω^Y are the medians

of the observations obtained for X and Y respectively. Another such metric would be $|\bar{y}-\bar{x}|$ where $\bar{x}$ is the observed average $\frac{x_1 + \ldots + x_m}{m}$ and $\bar{y}$ is the observed average $\frac{y_1 + \ldots + y_n}{n}$. The difficulty with these metrics is that (as with alternatives to the median metric) they are no longer well-defined on the composite model M_o. If, however, we go to a more restricted universe, such as the universe of all pairs of normal distributions with a certain fixed preassigned value of variance then a metric such as $|\bar{y}-\bar{x}|$ becomes useful. It proves to be well-defined on the composite model of identical pairs in this smaller universe, and it gives more powerful tests and smaller confidence intervals. We study this further in Chapter 17.

Ties in the WMW metric. If, in an observation $(x_1,\ldots,x_m,y_1,\ldots,y_n)$ for the WMW metric, there are an x_i and a y_j such that $y_j = x_i$, we say that a tie occurs for the pair (x_i,y_j). To calculate the metric for an observation which has some ties, we count each different tie as contributing 1/2 to the value of U_ω. This is equivalent, in the calculation of T_ω, to assigning to each member of each group of equal values in the over-all pooled ordering the average of the ranks of the positions that those members occupy. For example, if $(x_1,\ldots,x_m) = (2.1,5.2,6.3)$ and $(y_1,\ldots,y_m) = (5.2,5.2)$, then the overall pooled ordering is $(2.1,5.2,5.2,5.2,6.3)$, and corresponding assigned ranks are $(1,3,3,3,5)$. Comparing (x_i,y_j) pairs, we get $U_\omega = 2 + 1/2 + 1/2 = 3$, and adding ranks, we get $T_\omega = 3 + 3 = 6$. Since $\frac{n(n+1)}{2} = 3$, we see that the relationship $T_\omega = U_\omega + \frac{n(n+1)}{2}$ is preserved. It is easy

to show that this relationship always holds when ties are counted in this way.

When ties occur, the DLS may be calculated in the usual ways. If it is calculated by listing all possible sets Ω of positions for the Y values and then seeing what proportion have $|U_\Omega - \frac{mn}{2}| \geq |U_\omega - \frac{mn}{2}|$, the resulting DLS value is correct since, for continuous and identically distributed random variables, a tie indicates a comparison that would be settled by more accurate measurement and settled half the time in one direction and half the time in the other. If the DLS is calculated by normal approximation, ties can be shown to lead to values that are somewhat larger than the correct DLS value. Unless the number of ties is quite large, this difference is slight. In any case, if the approximated DLS falls below the critical level in a hypothesis test, then the correct DLS must fall below as well.

Confidence intervals in the WMW metric. Let $f(x)$ be a density function for a continuous random variable. For each fixed real number d, we define $f^d(x)$ to be the density function $f(x-d)$. Then f^d may be pictured as the density function f after it has been shifted d units in the positive direction. (Hence, for example, the median of f and the median of f^d must differ by d.) We say that f and g have the same shape if $g = f^d$ for some d. We now take, as our universe of models, the set of all pairs of distributions in which the two distributions have the same shape. Each pair (f,g) in the universe has a unique value d associated with it, where d is taken so that $g = f^d$. For each real number d, let M_d be the composite model

consisting of all pairs of the form (f,f^d). Then M_o is the composite model consisting of all pairs in which the two distributions are identical. Let Ω be an observation $(X_1,\ldots,X_m,Y_1,\ldots,Y_n)$. Define ω^d to be the observation $(X_1+d,\ldots,X_m+d,Y_1,\ldots,Y_n)$. Let $s^{WMW}(\Omega)$ be the WMW metric. Finally, for each d, define the metric $s_d^{WMW}(\Omega)$ by the equation:

$$s_d^{WMW}(\Omega) = s^{WMW}(\Omega^d).$$

The metric s_d^{WMW} is well-defined on M_d, since it yields the same DLS value for Ω under (f,f^d) in M_d that s^{WMW} yields for Ω^d under (f,f) in M_o, and we have already seen that s^{WMW} is well-defined on M_o.

Given an observation ω and a confidence level γ, we can now get a confidence interval for values of d as follows. The confidence interval will be those values of d such that ω yields a DLS $> \alpha$ under the metric s_d^{WMW} (where $\alpha = 1-\gamma$). This rather abstract statement can be put in more concrete terms as follows. We ask the question: to what positions can we shift the X values so that the Y values and X values will then give DLS $> \alpha$ under the WMW metric? The largest and smallest shifts will be the confidence limits on d.

Example. Consider the data given above on tutored and untutored verbal aptitude scores. Take as universe all pairs of distributions such that the two distributions in that pair have the same shape. We find a 95% confidence interval for d, the difference in position of the distributions for X and for Y, as follows. Recall that we observed

Y : 585,590,609,614,622,625,

X : 576,600,606.

If we shift the X values to the left, we find that for a shift of d = -16, we get,

X : 560,584,590.

For slightly larger (negative) shifts, we have $U_\omega = 17$ which gives $\underline{DLS} = \frac{4}{84} = 0.048$, and for slightly smaller (negative) shifts we get $U_\omega = 16$ with $\underline{DLS} = \frac{8}{84} = 0.10$. Hence the lower confidence limit for d is -16. Similarly, in a shift to the right we get, for a shift of 46,

X : 622,646,652.

Again, slightly larger (positive) shifts give $\underline{DLS} = 0.048$, while smaller shifts give $\underline{DLS} = 0.10$. Hence the upper confidence limit is 46. Therefore, the 95% confidence interval on d, the difference in position between the distribution for X and the distribution for Y, is [-6,46].

Power of the Mann-Whitney test. Consider the case of an observation with m = n = 4 where the entire ordering has the form XXXYXYYY. It is easy to calculate that the DLS of this observation under the WMW metric is 0.06. This observation can also be analyzed as a type γ contingency table with

assigned margins (4,4) and (4,4), if we classify the entire sample into a top half and a bottom half with respect to X or Y values. We would have

3	1
1	3

.

A small sample (hypergeometric) calculation gives the DLS of this table to be 0.49. Thus the WMW metric, which takes into account details of the relative ordering of all the individual observed values gives a much more discriminating result than the contingency table which merely counts the number of values of each kind in the upper and lower halves of the entire ordering. If we use each of these metrics in a hypothesis test, it follows that the WMW metric gives a more powerful test. The contingency table test, on the other hand, requires less information and is easier to carry out. Note that for the ordering XXXXYYYY, the information used by the two metrics is equivalent. Not surprisingly, each then gives the same DLS (= 0.03).

Inventing a metric. It may be helpful for a student, in approaching a new problem, to define an appropriate universe of models and then to seek to define a new metric that will be especially suited to the problem. Often, the student will find that the method which he or she develops in this way will already have been studied and used by other statisticians. The exercise of seeking to define a new metric can add to the student's own insight and understanding of the given problem. In seeking to define a metric, we look for: (a) a metric that represents the circumstances of the problem in a reasonable and

intuitive way; and (b) a metric for which DLS values are easy to calculate (either exactly or approximately). In the paragraphs which follow, we will explore several new statistical situations. We will see how each situation leads us to a natural and useful metric. In each case, the metric will turn out to be a metric which is already familiar to statisticians.

The Kruskal-Wallis metric. We used the WMW metric to measure how well an observation $(x_1,\ldots,x_m,y_1,\ldots,y_n)$, that included a group of independent values for X and a group of independent values for Y, confirmed that two random variables X and Y had the same distribution. We now turn to the case of k random variables and ask how well an observation that includes a group of independent values for each of the random variables confirms that all k variables have the same distribution. We use the case $k = 3$ to illustrate, and we call the random variables X, Y, and Z.

Let ω be the observation $(x_1,\ldots,x_m,y_1,\ldots,y_n,z_1,\ldots,z_q)$ where $(x_1,\ldots,x_m)$ are m independent values of X, $(y_1,\ldots,y_n)$ are n independent values of Y, and $(z_1,\ldots,z_q)$ are q independent values of Z. A model for our experiment will be a triple of continuous distributions, where the first is for X, the second is for Y, and the third is for Z. We let the composite model M_o consist of all models in which the three distributions are identical. We now define a metric that generalizes the Mann-Whitney calculation by ranks. Let $N = m + n + q$. The entire set of N observed values is arranged in order of size. Each value is assigned a rank number. Let R_1, R_2, R_3 be the sums of the

ranks for the values of X, Y, and Z respectively. (Ties are treated as in the WMW case.) Then $R = R_1 + R_2 + R_3 = \frac{N(N+1)}{2}$, and the average rank in the entire set must be $\frac{N+1}{2}$. Intuitively, we would be inclined to say that an observation Ω confirms M_o especially well if, for each of the three groups of observations, the average of the ranks appearing in that group is close to the same value $\frac{N+1}{2}$. Hence we could measure deviation from such confirmation by the natural formula

$$s^{KW}(\Omega) = m(\frac{R_1}{m} - \frac{N+1}{2})^2 + n(\frac{R_2}{n} - \frac{N+1}{2})^2 + q(\frac{R_3}{q} - \frac{N+1}{2})^2 ,$$

where we have weighted the squared deviation in each group by the number of values in that group. This metric is called the Kruskal-Wallis metric. Algebraic simplification leads to the formula

$$s^{KW}(\Omega) = \frac{R_1^2}{m} + \frac{R_2^2}{n} + \frac{R_3^2}{q} - \frac{N(N+1)^2}{4} .$$

For k random variables with $n_1, n_2, \ldots, n_k$ independent values observed in the k groups, the formula becomes

$$s^{KW}(\Omega) = \sum_i \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \qquad (\text{where } N = \sum_i n_i).$$

It is easy to show that the metric is well-defined on M_o (by a similar argument to that used for the WMW metric.)

Given an observation ω how hard is it to calculate a DLS value under this metric? One approach would be to use a

computer to list all possible rank orderings and then to find the proportion of these that give values of the metric $\geq s^{KW}(\omega)$. Another approach is to seek a simple and quick approximation method. In Chapter 19, we shall see that chi-square curves can be used in the following way when $n_1,\ldots,n_k$ are sufficiently large. Let H_ω be the result of multiplying $s^{KW}(\omega)$ by $\frac{12}{N(N+1)}$. Then

$$H_\omega = \frac{12}{N(N+1)} \sum_i \frac{R_i^2}{n_i} - 3(N+1) ,$$

and

$$\underline{DLS}(\omega) \approx C_d(H_\omega), \quad \text{where} \quad d = k-1.$$

(Here, by definition, $\underline{DLS}(\omega) = P_\mu(s^{KW}(\Omega) \geq s^{KW}(\omega))$ for any μ in M_o.)

Example. Assume that we observe values as follows:

X : 75,73,67,62

Y : 70,68,53,50

Z : 69,61,58,51.

The corresponding ranks are:

X : 12,11,7,6

Y : 10,8,3,1

Z : 9,5,4,2.

Therefore $R_1 = 36$, $R_2 = 22$, $R_3 = 20$. Hence

$s^{KW}(\omega) = \frac{(36)^2}{4} + \frac{(22)^2}{4} + \frac{(20)^2}{4} - \frac{12(13)^2}{4} = 38$, and

$H_\omega = \frac{12(38)}{12(13)} = 2.92$. Finally, $C_2(2.92) = 0.24$, and this is our desired approximate DLS. (By computer, we find the exact DLS in this case to be 0.252.)

It is easy to show that when $k = 2$, s^{KW} is equivalent to the WMW metric. Hence the KW metric is a direct generalization of the WMW metric.

Matched groups. A common and general technique in mathematical statistics is the use of matched groups in comparing random variables. We begin with an example, then give brief theoretical discussion, and finally introduce three natural and useful metrics. Matched groups are also sometimes called blocks.

Example. We wish to test whether two different strains of laboratory rat (strain A and strain B) have the same distribution of body weights. We realize, however, that within each strain, typical body weights will vary with an animal's age. More specifically, within each strain, the probability distribution of body weights at one age may be different in shape from the probability distribution of weights at another age. To carry out a statistical comparison, we match animals according to age, and then, within each age group (or block), we compare observations of strain A with observations of strain B. If we did not match according to age in this way, variation by age within each strain might keep us from seeing differences between the strains.

Theoretical discussion. What would it mean in the above example to say that there was no difference between strain A and strain B? It would mean that for each age t, the probability distribution of weights in strain A was identical with the probability distribution of weights in strain B. Let the random variable X_t be the body weight of a rat in strain A chosen at random from rats of age t. Let Y_t be the same for strain B. Let $f_t(x)$ be the probability density for the random variable X_t. Similarly, let $g_t(x)$ be the density function for Y_t. Hence, to say that there is no difference between A and B is to say that $f_t(x) = g_t(x)$ for all values of the parameter t. If we rewrite $f_t(x)$ as $f(x,t)$ and $g_t(x)$ as $g(x,t)$, we can call f and g parameterized densities. We take, as our universe of models, the set of all pairs of parameterized densities. As the null hypotheses, we take the composite model M_o consisting of all identical pairs of parameterized densities. (Identical means that $f(x,t) = g(x,t)$ for all x and all t.)

The metrics that we now introduce concern experiments where we attempt to compare values of X_t and Y_t by first choosing various different parameter values t and then observing values of X_t and Y_t for each of these parameter values. (Note that the above example considered only one parameter: age. In general, there may be several different parameters or other experimental conditions that we will wish to match at the same time. For the rats, these might include such factors as age, previous diet, and freedom from disease.)

The sign metric. Our first metric arises when we make an observation of independent matched pairs. Assume that we consider pairs of rats where, in each pair, the two rats have the same age and there is one rat of strain A and one rat of strain B. Assume that we get the following observation

A	50	100	133	270	780	340	290
B	45	90	120	220	690	347	250

where each column gives weights for a pair of rats of the same age. A natural metric, in this case, can be defined by assigning a + to each pair where the A value exceeds the B value and a - to each pair where the B value exceeds the A value. We get the result

$$+ + + + + - + .$$

Under the null hypothesis, the sign from each matched pair can be viewed as the result of a Bernoulli trial with + as success and with $p = 1/2$. Hence we can use the usual binomial metric. We define the metric

$$s^+(\Omega) = |\sigma - \tfrac{n}{2}|, \quad \text{where } \sigma \text{ is the number of } +\text{'s}$$

(In this case, σ is not standard deviation.)

and n is the total number of pairs. s^+ is called the sign metric. s^+ is well-defined for models in M_o, and DLS values can be calculated in the usual way (either directly from the binomial distribution with $p = 1/2$ or by normal approximation). For the observation in the above example, we get

$$s^+(\omega) = |6-3.5| = 2.5.$$

Hence, $DLS(\omega) = P(S^+(\Omega) \geq 2.5)$

$$= [\binom{7}{0} + \binom{7}{1} + \binom{7}{6} + \binom{7}{7}]\frac{1}{2^7}$$

$$= \frac{16}{128} = 0.13.$$

A tie, for the sign metric, is a pair of equal observations. A tie is counted as a Bernoulli trial with half a success, and contributes $\frac{1}{2}$ to the value of σ.

The signed-rank metric. Our second metric also arises when we make an observation of independent matched pairs. In it we take account not only of the sign but also of the size of the observed differences. We proceed as follows in the above example. First, we calculate the difference between A and B for each pair. We get

5, 10, 13, 50, 90, -7, 40.

Next, we take the absolute values of these differences. We get

5, 10, 13, 50, 90, 7, 40.

Then we assign ranks to these absolute values in the usual way. We get

1, 3, 4, 6, 7, 2, 5.

These are called the unsigned ranks. Finally, we insert algebraic signs from the original list of differences. This gives

$$1, 3, 4, 6, 7, -2, 5.$$

These are called the signed ranks. Let T_{Ω}^{+} be the sum of the positive signed ranks. In the example, we get

$$T_{\omega}^{+} = 26.$$

If n is the total number of pairs in the observation, then the total of the unsigned ranks must be $\frac{n(n+1)}{2}$. If the null hypothesis holds, we would expect about half of this total from positive ranks and half from negative ranks. It is therefore natural to define the following metric:

$$s^{WSR}(\Omega) = |T_{\Omega}^{+} - \frac{n(n+1)}{4}|.$$

This is called the Wilcoxon signed-rank metric. For the above example, we get $s^{WSR}(\omega) = |26-14| = 12$. We can make a computer calculation of the exact DLS for an observation ω under this metric by listing all possible (2^n) assignments of sign to the n ranks and then seeing the proportion of these which give a value of $s^{WSR}(\Omega)$ that is $\geq s^{WSR}(\omega)$. Such a calculation in the above example gives

$$\underline{DLS}(\omega) = 0.047.$$

In Chapter 19, we shall see that when n is sufficiently large, an approximate value may be obtained as follows

$$\underline{\text{DLS}}(\omega) \approx 1 - 2A(\zeta), \quad \text{where}$$

$$\zeta = \frac{|T_\omega^+ - \frac{1}{4}n(n+1)| - 1/2}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} .$$

In the above example, this approximation gives

$$\zeta = \frac{|26-14| - 1/2}{\sqrt{\frac{1}{24}7(8)(15)}} = \frac{11.5}{\sqrt{35}} = 1.94,$$

yielding $\underline{\text{DLS}}(\omega) \approx 0.052$.

Two kinds of ties can occur in calculating the WSR metric. First, there may be ties in the rank ordering of the absolute values of the differences and second, one or more of the absolute values may themselves be zero. Ties of the first kind are treated as in the WMW metric (by assigning to each member of a group of tied values the average of the ranks of the positions occupied by members of the group). Ties of the second kind are treated by assigning half of their unsigned rank total to T_ω^+. (This is the logically correct procedure. In practice, statisticians often omit ties of the second kind from consideration in the entire calculation. The difference in the result is usually slight.) As with the WMW metric, the existence of ties leads to DLS values, under the approximation calculation, that are slightly larger than correct.

Both the sign metric and the WSR metric may be used to establish a confidence interval from an observation in exactly the same way as the WMW metric is used to get a confidence interval. For this purpose, we take as our universe the set of all pairs (f,g) of parametrized densities such that $f(x,t) = g(x-d,t)$ for some real number d and all x and t. The confidence interval is then an interval of values of d.

The Friedman metric. Our third metric arises when we consider an observation of n matched k-tuples of values for k different (parameterized) random variables and take as null hypothesis that all k variables have the same parameterized density. The metric is a generalization of the sign metric. We give an example with $k = 3$ and $n = 4$. We consider three different strains of laboratory rat, strain A, strain B, and strain C. We observe the body weights of the rats in four matched triples, where, in each triple, all members have the same age. We get

A:	90	115	235	572
B:	71	95	250	575
C:	58	104	230	560

We now rank the observations within each triple. This gives, as ranks,

A:	3	3	2	2
B:	2	1	3	3
C:	1	2	1	1

We then sum the ranks for each strain. This gives $R_1 = 10$, $R_2 = 9$, and $R_3 = 5$ for A, B, and C respectively. In the general case, the total of all ranks in each k-tuple must be $\frac{k(k+1)}{2}$; hence the total of all ranks must be $\frac{nk(k+1)}{2}$. If the null hypothesis holds and the random variables have the same distributions, we expect the sum of the ranks for each random variable to be approximately the same. Hence, we expect the sum of the ranks for each random variable to be about $\frac{n(k+1)}{2}$. This suggests the natural metric

$$s^F(\Omega) = (R_1 - \frac{n(k+1)}{2})^2 + \ldots + (R_k - \frac{n(k+1)}{2})^2 .$$

This is known as the Friedman metric. Algebraic simplification leads directly to the formula

$$s^F(\Omega) = \sum_i R_i^2 - \frac{n^2k(k+1)^2}{4}.$$

As with the other metrics, a DLS value for ω can be got by a computer calculation of the proportion of all assignments of rank (there are $(k!)^n$ of them) that have values of $s^F(\Omega)$ at least as great as $s^F(\omega)$. In the above example we get $s^F(\omega) = 14$ and this gives $\underline{DLS}(\omega) = 0.273$. We shall see in Chapter 19 that when k is sufficiently large, an approximate calculation can be made using chi-square curves as follows. Let

$$F_\omega = \frac{12}{nk(k+1)} S^F(\omega).$$

Then, for an observation ω ,

$$\underline{DLS}(\omega) \approx C_d(F_\omega), \quad \text{where} \quad d = k-1.$$

In the above example, this gives

$$\underline{DLS}(\omega) \approx C_2(\tfrac{1}{4}14) = 0.18.$$

(The approximation to the exact DLS (0.27) is poor because k is small.) Ties in the Friedman metric are treated, within each k-tuple, in the same way as for the WMW metric.

Tables. Computer calculations of exact DLS values in certain specific cases for the WMW metric, the Kruskal-Wallis metric, the WSR metric, and the Friedman metric have been made and tabulated. The results may be found in published collections of statistical tables. (In the case of the WSR metric, the quantity usually used in the tables is T = the minimum of T_ω^+ and T_ω^-, where T_ω^- is the absolute value of the sum of the negative ranks.

Discrete random variables. Statisticians often apply the metrics described above to discrete random variables, and then go on to carry out DLS calculations exactly as above. They do so when they believe that the discrete distributions can be viewed as approximately continuous without much loss of accuracy (in the same sense that a binomial distribution, for large enough n, can be viewed as approximately a normal distribution.)

EXERCISES ON CHAPTER 16

16 - 1. In an experiment to test the rainmaking effectiveness of seeding clouds with iodide crystals, 13 storms were identified. Eight were selected at random and the clouds seeded. Average recorded rainfall then proved to be as follows

Treated storms: .06, .13, .15, .28, .41, .62, .83, 1.26

Untreated storms: .02, .09, .21, .29 1.09

State a null hypothesis. Use the WMW metric to calculate an approximate DLS, and hence decide if you would continue to accept at critical level 0.10.

16 - 2. A group of 29 students take an intelligence test. Before they do so, 14 are chosen at random to take a sample practice test. Results on the final test, taken by all students, are as follows.

Took sample test: 97, 108, 111, 112, 114, 118, 120, 121, 123, 125, 126, 128, 131, 139

No sample test: 94, 95, 98, 100, 101, 102, 105, 107, 108, 109, 113, 117, 119, 122, 127.

State a null hypothesis and use the WMW metric to calculate a DLS. (Remember to decide between one-sided and two-sided approach.)

16-3. In a survey of the period for which frozen orange juice was kept on the freezer shelf in a retail store, three brands were considered and eight cans of each brand were traced. The following observations (in days of storage) were obtained.

A	B	C
27	42	47
24	42	48
34	53	52
32	57	47
31	44	59
20	63	63
24	47	47
32	46	69

State a null hypothesis and use the KW metric to calculate a DLS.

16- 4. In the data from problem 3, use the Mann-Whitney test to test for a difference between brand B and brand C.

16- 5. Fifteen patients are used to compare two diuretics. Each patient is given one drug (chosen at random from the two under test) and then, after an interval of 6 days, the other. The observations obtained (litres of urine in 24 hours) were as follows:

Patient	Drug A	Drug B
1	1.66	2.24
2	2.01	2.18
3	1.84	2.40
4	0.62	1.30
5	2.25	2.57
6	1.17	1.87
7	1.20	1.38
8	1.04	1.58
9	2.50	2.79
10	2.39	3.16
11	1.04	1.53
12	1.55	2.19
13	3.90	4.61
14	2.11	2.67
15	1.76	1.56

State a null hypothesis and use the sign metric to get a DLS.

16- 6. In an experiment validating laboratory technique in a large hospital, 3 technicians made a certain measurement, each technician repeating the measurement with 5 different instruments. All 15 observations were made on subsamples from a single sample of material. Results were as follows.

	Instruments				
	1	2	3	4	5
Technician A	5.2	11.7	2.1	4.7	10.6
B	7.9	12.0	6.4	5.1	10.8
C	4.1	6.2	3.8	3.2	9.2

Use the Freidman metric twice: once to test that the technicians do not differ and once to test that the instruments do not differ.

16—7. 20 students are divided into 10 matched pairs on the basis of performance in a previous course. One number of each pair is selected at random and given TV instruction. The other attends a regular class. Results on the final are:

TV:	35	6	18	25	37	48	49	53	81	89
Regular:	34	57	9	64	61	75	91	100	98	93

State a null hypothesis. Calculate a DLS by the sign metric and then by the WSR metric.