4. STATISTICAL HYPOTHESIS TESTING

In all of the previous sections of these notes, we have focussed on the area of statistical estimation. In other words, we have tried to use our data (and sometimes a prior belief in the case of Bayesian approaches) to arrive at either "best guesses" (in the case of point estimation) or "plausible ranges" (in the case of interval estimation) for some quantitative aspect (often encoded as a parameter of a distributional family) of a population of interest. In many situations, however, the simple estimation of a population characteristic is not the final desired outcome of a statistical analysis. Specifically, we may want to use our estimates to decide whether some previously proposed theory or statement regarding the population of interest is actually true (or at least is plausible given the information provided by the observations at hand). This is, of course, the standard framework of statistical hypothesis testing which is familiar from any introductory unit in basic statistics. In this final section of these notes, we will briefly discuss the more formal structure of the theory of hypothesis testing which underlies the standard testing procedures (for population means and proportions) which are the main staple of any introductory presentation. We start by giving a formal set of definitions for parametric hypothesis testing and then introduce perhaps the most important and flexible of all testing procedures, based on the likelihood function. Finally, as we have done with both point and interval estimation previously, we will briefly investigate procedures for some standard situations which are not (as heavily) dependent on the parametric assumptions which will underly our initial discussions of statistical testing theory.

4.1. Definitions

We shall introduce and define the key aspects of a statistical hypothesis test through a rather simple example. Suppose that we have purchased a light-bulb based on its advertised claim that the mean lifetime of such bulbs is at least 1000 hours. If we then observe the lifetime of the actual bulb we purchased, we have some data with which to assess the advertising claim. This simple scenario is precisely the framework of statistical hypothesis testing.

4.1.1. Statistical Hypotheses and Decision Rules: More formally, suppose we believe that the lifetime of the population of bulbs in question is exponentially distributed with mean parameter θ, so that the probability density associated with X, the random lifetime of a bulb, is given by $f_X(x;\theta) = \theta^{-1}e^{-x/\theta}$ for some $\theta \in \Theta$. A *statistical hypothesis* is then simply a statement regarding the population of interest or, equivalently in the parametric case described here, the value of the true population parameter. As such, we can formulate the hypothesis we wish to examine regarding the population of light-bulbs as $H_0 : \theta \geq 1000$. More generally, we have:

> **Definition 4.1**: Suppose that $X_1, \ldots, X_n$ represent a simple random sample from a parametric family with density function $f_X(x;\theta)$ for some parameter $\theta \in \Theta$. A statistical hypothesis is simply a subset of the parameter space, Θ. Any statistical hypothesis of interest, often termed the *null hypothesis*, is associated with a competing *alternative hypothesis*. As such, a null hypothesis and its alternative form a partition of the parameter space Θ consisting of the sets: Θ_0, the set of parameter values which constitute the null hypothesis and $\Theta_1 = \Theta_0^c \cap \Theta$, the set of parameter values which are in the parameter space but not in the null hypothesis collection.

Note that in our light-bulb example, $\Theta_0 = \{\theta \in \Theta : \theta \geq 1000\}$. Moreover, we stress that the alternative hypothesis is defined as the complement of the null hypothesis *within the parameter space*. In other words, if we are considering testing the mean of a normal distribution and our null hypothesis is $H_0 : \mu = 0$, then the general alternative (in the case that the parameter space of μ is the entire real line) would be the *two-sided* one, $H_1 : \mu \neq 0$. However, if we restrict the parameter space to only non-negative values (perhaps because of some external information regarding the specific problem at hand), then the relevant alternative hypothesis would be the *one-sided* one, $H_1 : \mu > 0$, since $\{\mu = 0\}^c \cap \{\mu \geq 0\} = \{\mu > 0\}$.

A *statistical test* of the null hypothesis $H_0 : \theta \in \Theta_0$ is then just a *decision rule* based on the observed data for deciding whether to accept H_0 or reject it, and thus accept the alternative hypothesis, $H_1 : \theta \in \Theta_1$.

> **Definition 4.2**: Suppose that $X_1, \ldots, X_n$ represent a simple random sample from a parametric family with density functions $f_X(x;\theta)$ for some parameter $\theta \in \Theta$. Further let $\mathcal{X}$ represent the sample space of the (random) vector $X = (X_1, \ldots, X_n)$. A statistical test of the null hypothesis $H_0 : \theta \in \Theta_0$ is just a decision rule based on a partitioning of the sample space. In particular, if we partition the sample space $\mathcal{X}$ into those outcomes of the observations which would lead us to reject H_0, often denoted as C and referred to as the *rejection region* or *critical region* of the test, and those observations which would lead us to accept H_0, which is just the collection $C^c \cap \mathcal{X}$, then a statistical test is simply defined by the decision rule which rejects H_0 in favor of H_1 in the case that $X = (X_1, \ldots, X_n) \in C$ and accepts H_0 otherwise.

So, characterising a statistical test is as simple as defining its associated rejection region. For instance, in our light-bulb example, we can define the test which rejects H_0 if X, the observed lifetime of our sampled bulb, is less than 1000 hours. In other words, we define a test with critical region $C = \{X < 1000\}$. Indeed, since we have already seen (during our initial discussions of the concept of sufficiency) that statistics can be viewed as partitioning the sample space of the observations, it is quite common to define a statistical test in terms of a rejection region which is just a *level set* for some statistic $T(X_1, \ldots, X_n)$; in other words, C has the form $C = \{X \in \mathcal{X} : T(X) < k\}$ for some prespecified value k. Of course, whether this is a "good" test must be determined by examining the properties of the testing procedure so determined. This exercise is the subject of the next section.

4.1.2. Size and the Power Function: Common sense would indicate that the test described in the example of the previous section; namely, rejecting the null hypothesis that the mean lifetime of the bulbs is at least 1000 hours based on a single observation being less than 1000 hours, is not a very good test since it is quite prone to making an error. Indeed, we can assess the quality of a statistical test by examining the two distinct types of errors that can arise from it. If the observations fall in the rejection region C when in fact that null hypothesis, H_0, is true then our testing procedure will reject H_0 when it should not. Such a mistake is termed a *Type I error* and has a chance of occuring $Pr_\theta(C)$ for $\theta \in \Theta_0$. Alternatively, if the observed data values fall outside the rejection region when in fact the null hypothesis is false, then our testing procedure will accept H_0 when it should not. Such a mistake is termed a *Type II error* and has a chance of occuring $Pr_\theta(C^c)$ for $\theta \in \Theta_1$. Clearly, we would like to use a testing procedure which has a small chance of making errors of either type.

Of course, to actually assess the probability of making an error, we must make a probability statement about the observed data values, and these values depend on the true parameter θ. For instance, suppose that in our light-bulb example, the true mean lifetime of bulbs is exactly 1000 hours. In this case, H_0 is indeed true and the chance of a Type I error is

$$Pr(C) = Pr_{1000}(X < 1000) = \int_0^{1000} \frac{1}{1000} e^{-x/1000} dx = 1 - e^{-1000/1000} = 0.632.$$

Similarly, if $\theta = 1500$ the chance of a Type I error is

$$Pr(C) = Pr_{1500}(X < 1000) = \int_0^{1000} \frac{1}{1500} e^{-x/1500} dx = 1 - e^{-1000/1500} = 0.077.$$

On the other hand, if $\theta = 500$ then H_0 is false and the chance of making a Type II error is:

$$Pr(C^c) = Pr_{500}(X \geq 1000) = \int_{1000}^{\infty} \frac{1}{500} e^{-x/500} dx = e^{-1000/500} = 0.135.$$

Clearly, there is a strong relationship between Type I and Type II errors. In particular, note that for a given value of θ, only one type of error can occur (since for any given θ, H_0 either is or is not true). For convenience we generally focus our attention on the so-called *power function*:

Definition 4.3: The power function of a statistical test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ determined by the rejection region C is given by $K_C(\theta) = Pr_\theta(C)$. Note that this function yields the chance of a Type I error when $\theta \in \Theta_0$ and yields the probability of *correctly* rejecting H_0 when $\theta \in \Theta_1$; that is $K_C(\theta) = 1 - Pr_\theta(C^c)$, which is one minus the probability of a Type II error when $\theta \in \Theta_1$. This last probability is often termed the *power* of the test, since it represents the likelihood of the test detecting that the null hypothesis is indeed false (i.e., its *power of detection*).

Since $K_C(\theta)$ is just the chance of rejecting H_0 when the true parameter value is θ, we would like to have tests which have values of $K_C(\theta)$ which are large when $\theta \in \Theta_1$ and which are small when $\theta \in \Theta_0$. Of course, since $K_C(\theta)$ is a function, it can sometimes be difficult to work with directly. Therefore, we often define:

Definition 4.4: The *size* (or *significance level*) of a statistical test is given by

$$\alpha_C = \sup_{\theta \in \Theta_0} K_C(\theta).$$

In other words, the size of a test is the largest possible chance of a Type I error.

In the case of our light-bulb example, it is easy to calculate the power function as $K_C(\theta) = Pr_\theta(X < 1000) = 1 - e^{-1000/\theta}$. Therefore, the size of the test determined by $C = \{X < 1000\}$ is easily seen to be

$$\sup_{\theta \in \Theta_0} K_C(\theta) = \sup_{\theta \geq 1000} \left(1 - e^{-1000/\theta}\right) = 1 - e^{-1000/1000} = 0.632,$$

since $K_C(\theta)$ is clearly a decreasing function of θ in this case.

As we have noted, this seems a rather large chance of an error. Indeed, it is rather standard practice to focus on tests which have sizes on the order of 0.05 or 0.01. The problem with the current test is that it is too liberal with its rejection policy. With this in mind, we can define a new test based on a critical region which makes it more difficult to reject H_0. Suppose we define our new test based on the rejection region $C' = \{X < 250\}$. Some simple calculations (following along the lines of those set out for the critical region C, and left as an exercise for the reader) show that the power function for this new test is $K_{C'}(\theta) = 1 - e^{-250/\theta}$ which implies that the size of our new test is $1 - e^{-250/1000} = 0.221$. This value is certainly more palatable than that of the previous test, but it still seems rather large. In fact, as is commonly shown in introductory statistics units, we generally want to choose a rejection region to achieve a specific size, α. In other words, if we continue to focus on tests with rejection regions of the form $C_\alpha = \{X < k_\alpha\}$, we would like to choose k_α such that

$$\sup_{\theta \in \Theta_0} K_{C_\alpha}(\theta) = \sup_{\theta \geq 1000} \left(1 - e^{-k_\alpha/\theta}\right) = 1 - e^{-k_\alpha/1000} = \alpha.$$

Some simple algebra shows that, for this example, we have $k_\alpha = -1000 \ln(1 - \alpha)$. In particular, if we want a test with size $\alpha = 0.05$, we should use a test with rejection region $C = \{X < 51.29\}$, since $-1000 \ln(1 - 0.05) = 51.29$.

Of course, size (i.e., Type I error) is only one side of the coin. We must also examine our chance of a Type II error. In particular, we would like to ensure that the power function of our test is large when $\theta \in \Theta_1$. Note that if we employ the test based on the critical region $C = \{X < 51.29\}$ for our lightbulb example (so that we have a test with size $\alpha = 0.05$), then the power of this test when $\theta = 500$ (i.e., when the true mean lifetime is half of the advertised duration) is given by $K_C(500) = Pr_{500}(X < 51.29) = \int_0^{51.29} \frac{1}{500} e^{-x/500} dx = 1 - e^{-51.29/500} = 0.0975$. In other words, this test has less than a 10% chance of detecting even this drastic departure from the null hypothesis. Unfortunately, if our power is not as large as we like, then we cannot simply change a rejection region of the form $C = \{X < k\}$ to increase the power without simultaneously affecting the size of our test. Indeed, it is usually the case that simple modifications to a testing procedure to decrease the chance of a Type II error (or equivalently to increase the power of the test when H_0 is false) will increase the size of the test. Our task, then, is to find tests (or equivalently rejection regions) of a given size which have the best possible power when $\theta \in \Theta_1$. We note that there are two potential ways of modifying our test so as to increase its power at the same time as maintaining its size. The first is to change the sample space $\mathcal{X}$ (recall that a statistical test is equivalent to a partitioning of the sample space). The only way to effectively achieve a change in $\mathcal{X}$ is to change the sample size (and indeed, it should seem reasonable that the easiest way to increase the power of detection of departures from the null hypothesis is to increase the information available on which to base a decision). While this is sometimes a possibility in practice, usually we are in the position of already having gathered our observations and so the size of the sample is a fixed quantity. The other method of changing our test is, of course, to change our critical region C. We have noted that critical regions are generally based on level sets of a statistic (though they certainly do not have to be), and simply changing the level of the set [i.e., changing the value k in the region of the form $\{X \in \mathcal{X} : T(X) < k\}$] will generally only increase the power at the expense of increasing the size of the test as well. As such, we must change our critical region (and thus the corresponding test) more substantially and dramatically, generally by basing it on a different statistic, $T(X)$. Finding "good" tests based on level sets of statistics $T(X)$ for a fixed sample size is the subject of the following sections.

4.2. Most Powerful Tests

As noted at the end of the last section, we would like to determine tests which have a specified size and as large a power as possible when $\theta \in \Theta_1$. In fact, our goal will be to find a test which is *uniformly most powerful (UMP)* among all tests of a specified size α. Formally, we would like to find a test determined by a critical region C such that:

i. $K_C(\theta) \leq \alpha$ for all $\theta \in \Theta_0$; and,

ii. $K_C(\theta) \geq K_{C'}(\theta)$ for all $\theta \in \Theta_1$ and all subsets $C' \subseteq \mathcal{X}$ such that $K_{C'}(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.

A test (or critical region) satisfying this definition is a uniformly most powerful test of size α. Note that the second condition simply states that the test determined by the critical region C must have higher power at all points of Θ_1 than any other test (determined by critical region C') with size α. Unfortunately, as we have seen throughout our investigations, it is rarely possible to find such uniformly optimal procedures, since it will generally be the case that any given test will have the best power for some θ values in Θ_1 but not all. There are some special situations in which certain tests are known to be *UMP*. For example, if we assume a normal probability model, then the standard *one-sided* t-tests for differences in means and linear regression parameters can be shown to be uniformly most powerful (of course, if we assume some other, non-normal, probability model for our observations than this statement ceases to be true). Outside these cases, however, *UMP* tests rarely exist (and even if they do exist, actually finding them is usually extremely difficult).

However, there is one case in which it is always possible to find a *UMP* test, and this is the subject of the next section.

4.2.1. Simple Hypotheses and the Neyman-Pearson Lemma: A statistical hypothesis which consists of only a single parameter value is generally termed *simple*. For instance, if Θ_0 for the null hypothesis $H_0 : \theta \in \Theta_0$ consists of the single value θ_0 (i.e., $\Theta_0 = \{\theta_0\}$), then it is a simple hypothesis. Alternatively, if a hypothesis is not simple (i.e., it contains more than a single possible value) it is termed *composite*. In this section, we will examine the case of a statistical test for which both the null and alternative hypotheses are simple. In other words, we shall suppose that $X_1, \ldots, X_n$ are a sample from a population characterised by a probability model with density function $f_X(x;\theta)$ for $\theta \in \Theta$ where $\Theta = \{\theta_0, \theta_1\}$ and we shall focus on testing the null hypothesis $H_0 : \theta \in \Theta_0$ with $\Theta_0 = \{\theta_0\}$. Note that the structure of Θ means that this is a test of the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta = \theta_1$.

We now demonstrate that the *UMP* test in the case of two simple hypotheses is based on the so-called *likelihood ratio*:

$$\Lambda(X_1, \ldots, X_n) = \frac{L(\theta_0; X_1, \ldots, X_n)}{L(\theta_1; X_1, \ldots, X_n)} = \frac{f_X(X_1, \theta_0) \cdots f_X(X_n; \theta_0)}{f_X(X_1, \theta_1) \cdots f_X(X_n; \theta_1)}.$$

In particular, the test we shall define has a critical region of the form $C = \{\Lambda(X_1, \ldots, X_n) \leq k\}$. [NOTE: Since θ_0 and θ_1 are specified constants in the current testing framework, $\Lambda(X_1, \ldots, X_n)$ is a statistic.] The idea here is to construct the critical region, C, by collecting together those elements of the sample space, $\mathcal{X}$, which give the strongest evidence against the null hypothesis. In this respect, the ratio of the likelihood for any given sample at each of the two possible parameter values is precisely a relative measure of how plausible the two hypotheses are. In other words, when $\Lambda(X_1, \ldots, X_n)$ is very small, this is strong evidence that the observations arose from the alternative hypothesis rather than the null hypothesis. All that remains, then, is to determine the value of k so as to ensure that the test is of the desired size α. This can always be accomplished with an application (of perhaps rather tedious) calculus in the current setting since we have assumed that our hypotheses are simple (and thus completely determine the distribution of the data). Of course, while it should seem intuitively reasonable that the likelihood ratio is a good method of distinguishing between samples which support the null hypothesis versus samples which support the alternative hypothesis, in order to be assured that the test based on this statistic is *UMP* we need to demonstrate that the likelihood ratio provides the "best" information for making this distinction. This fact is the subject of the so-called *Neyman-Pearson Lemma*:

Theorem 4.1: Suppose that $X_1, \ldots, X_n$ are a sample from a population characterised by a probability model with density function $f_X(x;\theta)$ for $\theta \in \Theta$ where $\Theta = \{\theta_0, \theta_1\}$ and we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ with $\Theta_0 = \{\theta_0\}$. Then the test with critical region $C = \{\Lambda(X_1, \ldots, X_n) \leq k_\alpha\}$, where $\Lambda(X_1, \ldots, X_n)$ is the likelihood ratio statistic defined previously and k_α is defined such that $Pr_{\theta_0}(C) = \alpha$, is uniformly most powerful among all tests of size no larger than α.

Proof: We start by defining any other test of size $\alpha' \leq \alpha$, determined by the critical region C'. We need to show that $Pr_{\theta_1}(C) \geq Pr_{\theta_1}(C')$, since this demonstrates that the test based on the critical region C has larger power than any other test of size no larger than α for all $\theta \in \Theta_1$ (and here we see why the fact that the alternative hypothesis is simple makes this situation much easier to deal with than the general case of a composite alternative). Now, we note the following simple probability identities:

$$Pr_{\theta_1}(C) = Pr_{\theta_1}(C \cap C') + Pr_{\theta_1}(C \cap C'^c)$$
$$Pr_{\theta_1}(C') = Pr_{\theta_1}(C' \cap C) + Pr_{\theta_1}(C' \cap C^c),$$

which together imply that:

$$\begin{aligned} Pr_{\theta_1}(C) &= \{Pr_{\theta_1}(C') - Pr_{\theta_1}(C' \cap C^c)\} + Pr_{\theta_1}(C \cap C'^c) \\ &= Pr_{\theta_1}(C') + \{Pr_{\theta_1}(C \cap C'^c) - Pr_{\theta_1}(C' \cap C^c)\}. \end{aligned}$$

So, we can demonstrate the desired result by simply showing that $Pr_{\theta_1}(C \cap C'^c) - Pr_{\theta_1}(C' \cap C^c) \geq 0$. To do so, we first note that for any event $E \subseteq C$, we have:

$$\begin{aligned} Pr_{\theta_1}(E) &= \int_E L(\theta_1; x_1, \ldots, x_n) dx_1 \cdots dx_n = \int_E \frac{1}{\Lambda(x_1, \ldots, x_n)} L(\theta_0; x_1, \ldots, x_n) dx_1 \cdots dx_n \\ &\geq \frac{1}{k_\alpha} \int_E L(\theta_0; x_1, \ldots, x_n) dx_1 \cdots dx_n = \frac{1}{k_\alpha} Pr_{\theta_0}(E), \end{aligned}$$

since, by the definition of C, we have $\Lambda(x_1, \ldots, x_n) \leq k_\alpha$ for any $x = (x_1, \ldots, x_n) \in E \subseteq C$. Moreover, a nearly identical argument shows that for any event $F \subseteq C^c$ we have $Pr_{\theta_1}(F) \leq \frac{1}{k_\alpha} Pr_{\theta_0}(F)$. Therefore, since $C \cap C'^c \subseteq C$ and $C' \cap C^c \subseteq C^c$, we see that:

$$\begin{aligned} Pr_{\theta_1}(C \cap C'^c) - Pr_{\theta_1}(C' \cap C^c) &\geq \frac{1}{k_\alpha} Pr_{\theta_0}(C \cap C'^c) - \frac{1}{k_\alpha} Pr_{\theta_0}(C' \cap C^c) \\ &= \frac{1}{k_\alpha} \big\{ Pr_{\theta_0}(C \cap C'^c) - Pr_{\theta_0}(C' \cap C^c) \big\} \\ &= \frac{1}{k_\alpha} \big\{ Pr_{\theta_0}(C \cap C'^c) + Pr_{\theta_0}(C \cap C') \\ &\qquad\qquad - Pr_{\theta_0}(C' \cap C) - Pr_{\theta_0}(C' \cap C^c) \big\} \\ &= \frac{1}{k_\alpha} \big\{ Pr_{\theta_0}(C) - Pr_{\theta_0}(C') \big\} \\ &= \frac{1}{k_\alpha} (\alpha - \alpha') \\ &\geq 0. \end{aligned}$$

So, we now have a *UMP* test for the case of simple null and alternative hypotheses. Of course, for any specific instance, we will need to calculate the appropriate value of k_α.

Example 4.1: Suppose that $X_1, \ldots, X_n$ are a random sample from a normal distribution with mean μ and unit variance. Further, suppose that we know $\mu \in \{0, 1\}$. We wish to test $H_0 : \mu = 0$ versus $H_1 : \mu = 1$. Now, the likelihood function in this case is:

$$L(\mu; X_1, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}} \exp\bigg\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2 \bigg\}.$$

Therefore, the uniformly most powerful test of size α in this case is determined by the rejection region $C = \{\Lambda(X_1, \ldots, X_n) \leq k_\alpha\}$, where

$$\begin{aligned} \Lambda(X_1, \ldots, X_n) &= \frac{(2\pi)^{-n/2} \exp\big\{ -\frac{1}{2} \sum_{i=1}^{n} X_i^2 \big\}}{(2\pi)^{-n/2} \exp\big\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - 1)^2 \big\}} \\ &= \exp\bigg[-\frac{1}{2} \sum_{i=1}^{n} \{X_i^2 - (X_i - 1)^2\} \bigg] \\ &= \exp\bigg(\frac{n}{2} - \sum_{i=1}^{n} X_i \bigg), \end{aligned}$$

and k_α is determined so that $Pr_0(C) = \alpha$. Now, before we actually determine k_α directly, we note that the critical region determined by the likelihood ratio can often be simplified in structure. In this case, note that:

$$\{\Lambda(X_1,\ldots,X_n) \le k_\alpha\} = \left\{\exp\left(\frac{n}{2} - \sum_{i=1}^n X_i\right) \le k_\alpha\right\} = \left\{\frac{n}{2} - \sum_{i=1}^n X_i \le \ln(k_\alpha)\right\}$$
$$= \left\{\sum_{i=1}^n X_i \ge \frac{n}{2} - \ln(k_\alpha)\right\} = \left\{\overline{X} \ge \frac{1}{2} - \frac{1}{n}\ln(k_\alpha)\right\}.$$

In other words, we can now see that the *UMP* test is equivalently determined by a rejection region of the form $C = \{\overline{X} \ge c_\alpha\}$, where $c_\alpha = \frac{1}{2} - \frac{1}{n}\ln(k_\alpha)$ is now determined so that $Pr_0(C) = \alpha$. This form of the critical region makes determination of the required constant much easier, since the distribution of the statistic $\overline{X}$ is well-known in this case. In particular, when $\mu = 0$, $\overline{X}$ is normally distributed with mean 0 and variance $\frac{1}{n}$, so that the required value of c_α can be determined as:

$$\begin{aligned} Pr_0(\overline{X} \ge c_\alpha) = \alpha \qquad &\Longrightarrow \qquad Pr_0(\sqrt{n}\,\overline{X} \ge c_\alpha\sqrt{n}) = \alpha \\ &\Longrightarrow \qquad 1 - \Phi(c_\alpha\sqrt{n}) = \alpha \\ &\Longrightarrow \qquad c_\alpha = \Phi^{-1}(1-\alpha)\frac{1}{\sqrt{n}}. \end{aligned}$$

Of course, we can now determine the value of k_α if we so desire, but it is no longer necessary, as we see that the *UMP* test is now simply determined by the decision rule which rejects $H_0 : \mu = 0$ in favor of $H_1 : \mu = 1$ whenever $\overline{X} \ge \Phi^{-1}(1-\alpha)\frac{1}{\sqrt{n}}$. As a final aside, we note that this rejection rule can also be written in the form:

$$\frac{\overline{X} - 0}{\sqrt{1/n}} \ge \Phi^{-1}(1-\alpha),$$

which looks strikingly like the usual one-sided test for a single population mean when the population variance is assumed known. Indeed, this is precisely the starting point for demonstrating the previously stated facts regarding the *UMP* nature of the usual one-sided t-tests under the assumption of normally distributed observations.

We close this section with a few remarks. First, we note that the simple nature of the null and alternative hypotheses assumed here by no means requires θ_0 and θ_1 to be scalar values, just that they be a single (possibly vector-valued) point in the parameter space Θ. Second, we note that there was no real requirement that our observed sample contain independent observations or even that the structure of our testing framework be parametric in the true sense of the word. All that we truly required was that the two competing hypotheses each completely determined a distinct joint likelihood for the observed data. In other words, if our hypotheses took the form $H_0 : f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = g_0(x_1,\ldots,x_n)$ and $H_1 : f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = g_1(x_1,\ldots,x_n)$ where $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ represents the joint density function of the observations $X_1,\ldots,X_n$ and $g_0(x_1,\ldots,x_n)$ and $g_1(x_1,\ldots,x_n)$ are two given functions, then the *UMP* test of size α associated with these competing hypotheses is determined by a rejection region of the form $C = \left\{\frac{g_0(x_1,\ldots,x_n)}{g_1(x_1,\ldots,x_n)} \le k_\alpha\right\}$ where k_α is determined so that $\int_C g_0(x_1,\ldots,x_n)dx_1\cdots dx_n = \alpha$ (of course, this last requirement may mean a rather tedious and complicated calculus problem is required before we can actually implement this test). Finally, we note that it may not always be possible to find a value k_α which satisfies the strictures of Theorem 4.1. In other words, there may be no value k_α such that $Pr_{\theta_0}(C) = \alpha$ exactly. In particular, this can occur when the observed data have a discrete distribution.

Example 4.2: Suppose that $X_1, \ldots, X_{10}$ are *iid* random variables having a Bernoulli distribution with parameter θ. Further, suppose that we wish to test $H_0 : \theta = 0.5$ versus $H_1 : \theta = 0.2$. The likelihood function in this case is just:

$$L(\theta; X_1, \ldots, X_{10}) = \theta^{10\overline{X}}(1-\theta)^{10(1-\overline{X})},$$

where $10\overline{X} = \sum_{i=1}^{10} X_i$ is just the number of the X_i's which take the value 1. As such, the *UMP* test of a given size α is the one determined by the rejection region of the form:

$$\begin{aligned} C &= \{\Lambda(X_1, \ldots, X_n) \le k_\alpha\} \\ &= \left\{ \frac{0.5^{10\overline{X}} 0.5^{10(1-\overline{X})}}{0.2^{10\overline{X}} 0.8^{10(1-\overline{X})}} \le k_\alpha \right\} \\ &= \left\{ \left(\frac{5}{8}\right)^{10} 4^{10\overline{X}} \le k_\alpha \right\} \\ &= \{10\overline{X} \le c_\alpha\}, \end{aligned}$$

where $c_\alpha = \log_4 \left\{ \left(\frac{8}{5}\right)^{10} k_\alpha \right\}$. Suppose that we wish to find a *UMP* test of size $\alpha = 0.01$. Since $10\overline{X}$ has a binomial distribution with parameters $n = 10$ and $p = 0.5$ under the null hypothesis, we see that we must find c_α such that:

$$\sum_{i=0}^{c_\alpha} \binom{10}{i} 0.5^{10} = 0.01.$$

However, some simple arithmetic shows that:

$$\sum_{i=0}^{0} \binom{10}{i} 0.5^{10} = 0.0009765625; \qquad \sum_{i=0}^{1} \binom{10}{i} 0.5^{10} = 0.0107421875.$$

Therefore, $Pr_{0.5}(C) < 0.01$ for any choice $c_\alpha < 1$ and $Pr_{0.5}(C) > 0.01$ for any choice $c_\alpha \ge 1$. In other words, there is no possible value of c_α which makes the probability of the rejection region exactly equal to 0.01.

In such cases, however, while there may be no *UMP* test of a specific size α (if k_α does not exist for this size), there will always be a *UMP* test for some collection of sizes $\alpha_1, \alpha_2, \ldots$, and we can then pick the *UMP* test with the size closest to our desired size α. Indeed, in Example 4.2, we can find a *UMP* test of size $\alpha = 0.0107421875$ which is rather close to 0.01.

4.2.2. Generalised Likelihood Ratio Tests: In the previous section, we were able to find a *UMP* test in the case of simple null and alternative hypotheses. Of course, we have already noted that such an endeavour is generally not possible in the case of composite hypotheses. Nonetheless, the result of the Neyman-Pearson lemma does lead quite naturally to the construction of a test in the case of composite hypotheses. In particular, suppose that $X_1, \ldots, X_n$ are a random sample from a population characterised by a probability model with density function $f_X(x; \theta)$ for $\theta \in \Theta$ and we are interested in testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_0 \cup \Theta_1 = \Theta$ is any partition of the parameter space. Following the general notion of the Neyman-Pearson lemma, we can define the *generalised likelihood ratio*:

$$\Lambda_g(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta; X_1, \ldots, X_n)}{\sup_{\theta \in \Theta} L(\theta; X_1, \ldots, X_n)} = \frac{\sup_{\theta \in \Theta_0} f_X(X_1; \theta) \cdots f_X(X_n, \theta)}{\sup_{\theta \in \Theta} f_X(X_1; \theta) \cdots f_X(X_n, \theta)}$$

and the generalised likelihood ratio test which has critical region $C = \{\Lambda_g(X_1, \ldots, X_n) \le k_\alpha\}$ where, as usual, k_α is defined so that the size of the test is α; that is, $\sup_{\theta \in \Theta_0} Pr_\theta(C) = \alpha$. Note that the generalised likelihood ratio differs from the likelihood ratio statistic defined in Theorem 4.1 not only in the use of supremums (which are now necessary due to the potentially composite nature of the hypotheses) but also in that the denominator is maximised over the entirety of the parameter space Θ (rather than over the alternative hypothesis). This difference is employed for purely mathematical reasons, and a little thought shows that the set C defined by a level set of the generalised likelihood ratio is typically equivalent to a level set of the statistic:

$$\Lambda_g'(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta; X_1, \ldots, X_n)}{\sup_{\theta \in \Theta_1} L(\theta; X_1, \ldots, X_n)}.$$

In other words,

$$\{\Lambda_g(X_1, \ldots, X_n) \leq k_\alpha\} = \{\Lambda_g'(X_1, \ldots, X_n) \leq k_\alpha'\}$$

for some value k_α', provided the level set of $\Lambda_g(X_1, \ldots, X_n)$ in question has $k_\alpha < 1$. However, from the perspective of constructing a critical region, these are the only level sets of interest, since samples for which $\Lambda_g(X_1, \ldots, X_n) = 1$ indicate that the null hypothesis is at least as likely as the alternative (since the supremum of the likelihood over the entire space is no larger than the supremum over the null hypothesis subset) and as such would never reasonably be included in a rejection region.

It would be nice if this test based on the generalised likelihood ratio was always the *UMP* test, however, this is not the case. There are indeed cases where this test can be shown to be the *UMP* test (indeed, the usual t-tests for population means and linear regression coefficients in the case of normally distributed observations turn out to have the form of generalised likelihood ratio tests). However, a full demonstration of when these tests are *UMP* is beyond the scope of these notes. Moreover, even were we able to conclude that the likelihood ratio test was *UMP* we would still be in the unenviable position of having to determine the appropriate value k_α in the definition of the critical region C. Fortunately, it turns out that even when the generalised likelihood ratio test is not *UMP*, it typically has excellent properties (in particular, it can be shown to have nearly the largest possible power as the sample size increases towards infinity). As such, we tend to use the generalised likelihood ratio test in most complex testing situations where no other specific *UMP* test is available.

We close this section by noting one other strength of the generalised likelihood ratio test. Recall that we must determine the value of k_α in the definition of the rejection region $C = \{\Lambda_g(X_1, \ldots, X_n) \leq k_\alpha\}$. To do so requires the distribution of the statistic $\Lambda_g(X_1, \ldots, X_n)$ which can be quite complicated in general. However, in some specific situations, the distribution of $\Lambda_g(X_1, \ldots, X_n)$ can be accurately approximated. In particular, suppose that $\theta = (\theta_1, \ldots, \theta_p)$, so that the probability model parameter is a p-vector. Further suppose that Θ is an open subset of p-dimensional Euclidean space (for example, the entire p-dimensional Euclidean space itself or perhaps the positive quadrant, so that $\Theta = \{\theta \in I\!R^p : \theta_1 > 0, \ldots, \theta_p > 0\}$) and the null hypothesis we are interested in testing has the form $H_0 : \theta_1 = \theta_{1,0}, \ldots, \theta_q = \theta_{q,0}$ for some $q \leq p$. In this case, it can be shown that the distribution of $-2\ln\{\Lambda_g(X_1, \ldots, X_n)\}$ is approximately chi-squared with q degrees of freedom. As such, we can construct a test based on the generalised likelihood ratio with an approximate size α which is determined by the rejection region $C = \{-2\ln[\Lambda_g(X_1, \ldots, X_n)] \geq \chi_q^2(1-\alpha)\} = \{\Lambda_g(X_1, \ldots, X_n) \leq -0.5\exp[\chi_q^2(1-\alpha)]\}$.

Example 4.3: Suppose that we observe the array of independent random variables X_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ where X_{ij} is normally distributed with mean μ_i and variance σ_i^2 (i.e., a standard balanced one-way analysis of variance dataset). An important assumption for the validity of standard ANOVA procedures is that of homoscedasticity. Suppose we wish to test this assumption; that is, we wish to test the hypothesis $H_0 : \sigma_1^2 = \cdots = \sigma_I^2$. Note that this hypothesis is not quite in the form required for our chi-squared approximation to the generalised likelihood ratio test. However, a simple reparameterisation from $\sigma_1^2, \dots, \sigma_I^2$ to $\sigma_1^2, \tau_2^2 = \sigma_2^2 - \sigma_1^2, \dots, \tau_I^2 = \sigma_I^2 - \sigma_1^2$ shows that the null hypothesis can be written in the form $H_0 : \tau_2^2 = 0, \dots, \tau_I^2 = 0$. Now, the likelihood for this situation can readily be calculated as:

$$\begin{aligned}
&L(\mu_1, \dots, \mu_I, \sigma_1^2, \tau_2^2 \dots, \tau_I^2; X_{11}, \dots, X_{IJ}) \\
&\qquad = \frac{1}{(2\pi)^{IJ/2}} \left\{ \frac{1}{\sigma_1^2(\tau_2^2 + \sigma_1^2) \cdots (\tau_I^2 + \sigma_1^2)} \right\}^{J/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(X_{ij} - \mu_i)^2}{\tau_i^2 + \sigma_1^2} \right\},
\end{aligned}$$

where we have defined $\tau_1^2 = 0$. Some straightforward (though tedious) calculus shows that this likelihood is maximised at:

$$\hat{\mu}_i = \frac{1}{J} \sum_{j=1}^{J} X_{ij} = \overline{X}_i; \qquad \sigma_1^2 = \frac{1}{J} \sum_{j=1}^{J} (X_{1j} - \overline{X}_1)^2; \qquad \tau_i^2 = \frac{1}{J} \sum_{j=1}^{J} (X_{ij} - \overline{X}_i)^2 - \hat{\sigma}_1^2,$$

while under the null hypothesis the likelihood is maximised at:

$$\hat{\mu}_{i,0} = \frac{1}{J} \sum_{j=1}^{J} X_{ij} = \overline{X}_i; \qquad \hat{\sigma}_{1,0}^2 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{1j} - \overline{X}_1)^2.$$

Substituting these values into the likelihood then yields:

$$\begin{aligned}
\sup_{\theta \in \Theta} L(\theta; X) &= \frac{1}{(2\pi)^{IJ/2}} \left\{ \frac{1}{\hat{\sigma}_1^2(\hat{\tau}_2^2 + \hat{\sigma}_1^2) \cdots (\hat{\tau}_I^2 + \hat{\sigma}_1^2)} \right\}^{J/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(X_{ij} - \hat{\mu}_i)^2}{\hat{\tau}_i^2 + \hat{\sigma}_1^2} \right\} \\
&= \frac{1}{(2\pi)^{IJ/2}} \left\{ \frac{1}{\hat{\sigma}_1^2(\hat{\tau}_2^2 + \hat{\sigma}_1^2) \cdots (\hat{\tau}_I^2 + \hat{\sigma}_1^2)} \right\}^{J/2} \exp\left(-\frac{IJ}{2} \right) \\
\sup_{\theta \in \Theta_0} L(\theta; X) &= \frac{1}{(2\pi)^{IJ/2}} \left(\frac{1}{\hat{\sigma}_{1,0}^2} \right)^{IJ/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(X_{ij} - \hat{\mu}_{i,0})^2}{\hat{\sigma}_{1,0}^2} \right\} \\
&= \frac{1}{(2\pi)^{IJ/2}} \left(\frac{1}{\hat{\sigma}_{1,0}^2} \right)^{IJ/2} \exp\left(-\frac{IJ}{2} \right).
\end{aligned}$$

Therefore, the generalised likelihood ratio statistic is given by:

$$\Lambda_g(X_{11}, \dots, X_{IJ}) = \frac{\sup_{\theta \in \Theta_0} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta; X)} = \left\{ \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_{1,0}^2} \right) \left(\frac{\hat{\tau}_2^2 + \hat{\sigma}_1^2}{\hat{\sigma}_{1,0}^2} \right) \cdots \left(\frac{\hat{\tau}_I^2 + \hat{\sigma}_1^2}{\hat{\sigma}_{1,0}^2} \right) \right\}^{J/2}.$$

Finally, then, we see that the generalised likelihood ratio test with approximate size α is determined by the rejection region:

$$\begin{aligned}
C &= \{-2\ln[\Lambda_g(X_{11}, \dots, X_{IJ})] \ge \chi_{I-1}^2(1-\alpha)\} \\
&= \left\{ IJ \ln(\hat{\sigma}_{1,0}^2) - J \sum_{i=1}^{n} \ln(\hat{\sigma}_1^2 + \hat{\tau}_i^2) \ge \chi_{I-1}^2(1-\alpha) \right\}.
\end{aligned}$$

We close this section by noting that a proof of the chi-squared approximation to the distribution of $-2\ln\{\Lambda(X_1, \dots, X_n)\}$ is beyond the scope of these notes, but it does follow along the lines of the argument used in the construction of the asymptotic chi-squared likelihood-based confidence intervals developed in Section 3.2.2.

4.3 Non-parametric Tests

To complete these notes, we present several hypothesis testing procedures for some standard situations which are not dependent on a proper choice of underlying parametric probability model for the observations. As such, these procedures are given the name *non-parametric*. This section is not intended to be a complete or rigorous development of the field of non-parametric tests, but rather a presentation of a few simple and common procedures which give the general flavour of the ideas involved in constructing non-parametric tests.

4.3.1. Tests for Univariate Samples: In this section, we shall assume that we have observed a random sample $X_1, \ldots, X_n$ from some univariate distribution having *CDF* $F(x)$. However, we shall not make any further assumptions regarding the form of $F(x)$. We shall then be interested in testing whether a specified quantile of the distribution is equal to some value. In other words, for given values p_0 and z, we wish to test the null hypothesis $H_0 : F(z) = p_0$ against the two-sided alternative $H_1 : F(z) \neq p_0$. One possible test in this case is to determine a rejection region based on a level set of the statistic $Y = \sum_{i=1}^{n} I_{\{X_i \le z\}}$, the number of sampled observations less than or equal to the null hypothesis value z. Suppose we define a critical region of the form $C = \{Y \le c_1 \textit{ or } Y \ge c_2\}$. The power function for this test is given by:

$$K_C(p) = Pr_p(C) = \sum_{i=0}^{c_1} Pr_p(Y=i) + \sum_{i=c_2}^{n} Pr_p(Y=i) = \sum_{i=0}^{c_1} \binom{n}{i} p^i (1-p)^{n-i} + \sum_{i=c_2}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

where p is the (unknown) true value of $F(z)$. Of course, we must choose c_1 and c_2 in order to achieve a desired size for this test. As such, we need to choose the values of c_1 and c_2 so that $K_C(p_0) = \alpha$. [NOTE: Since Y is clearly a discrete random variable, we will not be able to achieve all possible sizes; see Example 4.2 and the remarks at the end of Section 4.2.1.] We note that this test is valid regardless of the underlying distribution $F(x)$. Typically, the value of interest for p_0 will be one-half, so that we are testing whether z is the median of the distribution $F(x)$. In such cases, the test described here is referred to as the *sign test*, since it can be seen to be based on the number of positive values among the collection $X_1 - z, \ldots, X_n - z$. [NOTE: The version of the sign test presented here is two-sided, however, it can be easily modified to achieve a test against either of the one-sided alternatives $H_1 : F(z) > p_0$ or $H_1 : F(z) < p_0$. All that is required is a modification of the critical region to the form $C = \{Y \ge c\}$ or $C = \{Y \le c\}$, respectively.]

Example 4.4: Let $X_1, \ldots, X_{10}$ be a random sample from a population characterised by a distribution with *CDF* $F(x)$. Suppose we wish to test whether the median of this distribution is equal to 72; that is, we wish to test $H_0 : F(72) = 0.5$ against the two-sided alternative. Further, suppose that we would like a test of size $\alpha = 0.07$. Some simple calculation shows:

$$\sum_{i=0}^{0} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.00097656; \qquad \sum_{i=0}^{1} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.01074219;$$

$$\sum_{i=0}^{2} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.05468750; \qquad \sum_{i=8}^{10} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.05468750;$$

$$\sum_{i=9}^{10} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.01074219; \qquad \sum_{i=10}^{10} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.00097656.$$

So, we see that it is not possible to choose a rejection region such that the size of the test is precisely 0.07. However, we can choose either $C = \{Y \le 2 \textit{ or } Y \ge 9\}$ or $C = \{Y \le 1 \textit{ or } Y \ge 8\}$ and arrive at a test which has size $0.0547 + 0.0107 = 0.0654$ which is reasonably close to the desired level.

The sign test is remarkably flexible, making essentially no assumptions regarding the underlying distribution $F(x)$. However, it can be shown that its power (i.e., the probability of it detecting that the null hypothesis is actually false) is quite low (and indeed, calculating the power when $F(z) = p_1$ for some value $p_1 \neq p_0$ is a straightforward calculation again involving the binomial distribution). This lack of power should not be very surprising, as the sign test is only based on whether the given observations are larger than the proposed median z, and ignores how far above or below z the observations were. As such, the sign test tends to ignore useful information contained in the sample. It does this in order to avoid making various assumptions regarding the underlying distribution $F(x)$, and this is a common theme for non-parametric tests; namely, they give up power in order to avoid making parametric assumptions. This is not an advisable thing to do if we truly believe in a given set of parametric assumptions. However, if we do not believe in any parametric framework, then using the non-parametric approach seems a more prudent way to proceed. Nonetheless, the loss of information inherent in the sign test seems rather dramatic, and it can often be improved upon without requiring parametric assumptions to be made.

We saw that the sign test for the median was based on the statistic Y, the number of values in the collection $X_1 - z, \ldots, X_n - z$ which were positive. This approach essentially ignores the size of the deviation between the observation and the proposed null hypothesis median value z. It turns out that it is possible to retain some of the information contained in the size of these differences without reverting to a parametric approach. In particular, suppose we define the quantities $Z_i = |X_i - z|$ and let R_i be the rank of Z_i in an ordered list of the values $Z_1, \ldots, Z_n$. For example, if $n = 3$ and $Z_2 < Z_3 < Z_1$, then we would have $R_1 = 3$ (since Z_1 is the largest of the Z_i's) while $R_2 = 1$ and $R_3 = 2$. Finally, we define

$$s_i = \begin{cases} -1 & \text{if } X_i < z \\ 0 & \text{if } X_i = z \\ 1 & \text{if } X_i > z \end{cases} .$$

A test of the null hypothesis $H_0 : F(z) = 0.5$ can then be constructed based on the level sets of the so-called *Wilcoxon signed-rank statistic*, $W = \sum_{i=1}^{n} s_i R_i$. The idea is that if z truly is the median of the population, then the ranks R_i will be evenly dispersed among the positive and negative $X_i - z$ values, and thus the statistic W will tend to be near zero. On the other hand, if z is not the true median, then the large deviations from z will tend to congregate on one side of z or the other, meaning that more of the large ranks will go with either the positive $X_i - z$ values (if the true median is larger than z) or the negative $X_i - z$ values (if the true median is smaller than z). In either case, the value of the statistic W will tend to be far from zero (in either direction). Therefore, we can construct a test against the two-sided alternative with rejection region $C = \{W \leq c_1 \text{ or } W \geq c_2\}$. Again, of course, we must determine the values c_1 and c_2 so as to ensure that the size of our test is the desired value, α (and again, there are the obvious one-sided versions of this test). Unfortunately, unlike the sign test, the distribution of the statistic W is no longer as simple as the binomial distribution of Y. Nonetheless, the distribution of W under H_0 can indeed be computed directly (and tables of its distribution for small sample sizes exist). Moreover, it can further be shown that the distribution of W under H_0 is approximately normal with mean zero and variance $\sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1)$ when the sample size n is large. The demonstration of this fact is beyond the scope of these notes, however, we do note that $W = \sum_{i=1}^{n} s_i R_i$ has the form of a sum, and thus it is not overly surprising that its distribution can be approximated by a normal distribution.

Example 4.5: Suppose that we observe the following 20 data values:

94.1, 93.3, 91.2, 93.0, 104.8, 100.6, 110.4, 94.1, 95.2, 102.1,
92.9, 102.7, 111.5, 88.4, 88.7, 105.0, 94.0, 99.1, 109.5, 97.3,

and we wish to construct an $\alpha = 0.01$ level test of $H_0 : F(95) = 0.5$ versus the two-sided alternative. So, the desired critical region has the form $C = \{W \leq c_1 \text{ } or \text{ } W \geq c_2\}$, and we need to choose c_1 and c_2 so that $Pr_{H_0}(C) = 0.01$ (at least approximately). Using the fact that, under H_0, W is approximately normally distributed with mean zero and variance $\frac{1}{6}20(21)(41) = 2870$, we see that

$$Pr_{H_0}(C) \approx \Phi\left(\frac{c_1}{\sqrt{2870}}\right) + \left\{1 - \Phi\left(\frac{c_2}{\sqrt{2870}}\right)\right\},$$

and thus choosing $c_2 = -c_1 = 2.575\sqrt{2870} = 137.95$ yields a test with size:

$$Pr_{H_0}(C) \approx \Phi(-2.575) + \{1 - \Phi(2.575)\} = 0.01.$$

[NOTE: These are certainly not the only possible choices for c_1 and c_2, but the symmetry of the resulting rejection region seems a sensible feature.] Now, to actually implement the test on the given data, we note that the $X_i - 95$ values are:

$$\begin{array}{cccccccccc} -0.9, & -1.7, & -3.8, & -2.0, & 9.8, & 5.6, & 15.4, & -0.9, & 0.2. & 7.1, \\ -2.1, & 7.7, & 16.5, & -6.6, & -6.3, & 10.0, & -1.0, & 4.1, & 14.5, & 2.3. \end{array}$$

The ranks of the absolute values of this collection are:

$$\begin{array}{cccccccccc} 2.5, & 5, & 9, & 6, & 16, & 11, & 19, & 2.5, & 1, & 14, \\ 7, & 15, & 20, & 13, & 12, & 17, & 4, & 10, & 18, & 8. \end{array}$$

[NOTE: In the case of tied values, we simply assign the average rank; for example, the two absolute values of 0.9 are the second and third smallest, so each is assigned a rank of 2.5. It should be noted, however, that if there are a large number of tied observations, the normal approximation to the distribution of W can become poor, and the procedure described here would need to be modified.] So, we can now calculate the Wilcoxon signed-rank statistic as:

$$\begin{aligned} W &= -2.5 - 5 - 9 - 6 + 16 + 11 + 19 - 2.5 + 1 + 14 \\ &\quad - 7 + 15 + 20 - 13 - 12 + 17 - 4 + 10 + 18 + 8 \\ &= 88. \end{aligned}$$

Since $88 \notin C$, we do not reject the null hypothesis. Of course, if we were to change the size of our test to $\alpha = 0.1$, then the rejection region would need to change accordingly. A simple calculation (left as an exercise) shows that (assuming we wish to maintain the symmetric aspect of our rejection region), the new critical region is given by $C = \{W \leq -88.13 \text{ } or \text{ } W \geq 88.13\}$. Again, we see that $88 \notin C$, but this time it is a very near thing. Indeed, we recall from our introductory units in statistics, the *p-value* of a testing procedure is the smallest size α for which the observed data falls in the rejection region. As such, we see that the p-value associated with this Wilcoxon signed-rank test for the observed data is very near to 0.1.

We close this section by noting that the Wilcoxon signed-rank test generally has much better power than the sign test and still does not require parametric assumptions. Of course, the power of the Wilcoxon signed-rank test is still generally less than that of parametric procedures, provided we believe that the required parametric assumptions are indeed true.

4.3.2. Tests for Bivariate Samples: In this section, we shall assume that we have observed two random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two independent, univariate populations characterised by distributions having *CDF*s $F(x)$ and $G(y)$, respectively. As in the previous section, we shall not make any further assumptions regarding the forms of $F(x)$ or $G(y)$. We shall then be interested in testing whether the two populations are characterised by the same distribution. In other words, we wish to test the null hypothesis $H_0 : F(z) = G(z)$ for all z against the two-sided alternative $H_1 : F(z) \neq G(z)$ for some z. We shall discuss several different tests for this situation.

The first test we shall discuss is essentially just a test for the equality of medians (and as such is usually referred to as the *median test*). The idea of the test is that if the two populations have the same distribution (or indeed, just the same median) than when the two samples are combined and the median of this combined collection calculated, we should expect half of each sample to fall below the combined median. Specifically, then, we define $Z = \text{median}\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$ and

$$V = \sum_{i=1}^{n} I_{\{X_i < Z\}},$$

so that V is just the number of X_i's which fall below the combined median Z. Clearly, if H_0 is true then we would expect V to be equal to $n/2$, and thus we shall construct our test with a rejection region of the form $C = \{|V - \frac{n}{2}| \geq k\}$. All that remains, is to determine the value of k to achieve a desired size α for our test. If we assume that the *CDF*s $F(x)$ and $G(y)$ are continuous, so that the chance of any of the X_i's or Y_j's being equal is zero (i.e., there is no chance of any ties in the combined collection of observations), then it can easily be seen that in order for $V = v$, we must choose v out of the n X_i's and $\frac{m+n}{2} - v$ of the m Y_j's to be less than Z. This is precisely the structure of the so-called *hypergeometric distribution*. In other words, we have:

$$Pr_{H_0}(V = v) = \frac{\begin{pmatrix} n \\ v \end{pmatrix} \begin{pmatrix} m \\ 0.5(m+n) - v \end{pmatrix}}{\begin{pmatrix} m+n \\ 0.5(m+n) \end{pmatrix}},$$

where we must be careful to interpret $\begin{pmatrix} m \\ 0.5(m+n) - v \end{pmatrix}$ to be zero when $0.5(m+n) - v < 0$. [NOTE: In the case that $0.5(m+n)$ is not an integer then, by convention, we simply use $0.5(m+n-1)$ instead, the idea being that we have thus ignored the observed value equal to Z, the combined median, which will always exist in a combined sample of odd size.] Now, if m and n are small enough, an exact calculation of the hypergeometric probabilities can be performed and an appropriate value for k can then be chosen to yield a test of the desired size. When m and n are large, however, such calculations are extremely time consuming. As such, we can approximate the distribution of V with a normal distribution when m and n are large (in this particular case, the normal approximation is quite accurate as soon as $m, n > 10$). It is a reasonably straightforward exercise to show that:

$$E_{H_0}(V) = \frac{n}{2}; \qquad Var_{H_0}(V) = \frac{mn}{4(m+n-1)} = \sigma_V^2.$$

Therefore, if we want a test of approximate size α, we should choose k such that:

$$Pr_{H_0}\left(\left|V - \frac{n}{2}\right| \geq k\right) \approx \Phi\left(-\frac{k}{\sigma_V}\right) + \left\{1 - \Phi\left(\frac{k}{\sigma_V}\right)\right\} = \alpha.$$

A simple calculation then shows that we should choose $k = \Phi^{-1}(1 - \alpha/2)\sigma_V$. As a specific example, suppose that we have $m = n = 20$. In this case, $\sigma_V^2 = \frac{400}{4(39)} = 2.5641$. Thus, a test with size $\alpha = 0.05$ would reject H_0 whenever V differed from $\frac{n}{2} = 10$ by more than $k = 1.96\sqrt{2.5641} = 3.14$; that is, we will reject H_0 if more than 13 or fewer than 7 X_i's fall below the combined median, Z.

Of course, just as the sign test ignored the actual values of the observed data, the median test described above does not take into account the size of the X_i's and Y_j's but only the number of these values which fall below the observed median of the combined sample. In the case of the univariate framework, we saw that the sign test could be improved upon by incorporating the ranks of the data, and this led to the Wilcoxon signed-rank test. In the current setting, a very similar approach can be taken to develop an improved test for the null hypothesis $H_0 : F(z) = G(z)$ for all z. We again start by considering the combined sample, and define R_i $(i = 1, \ldots, n)$ to be the rank of X_i in the combined sample. For example, if we have observed the samples $X_1 = 1, X_2 = 6, X_3 = 2$ and $Y_1 = 0, Y_2 = 4$, then the ordered combined collection is Y_1, X_1, X_3, Y_2, X_2 and thus $R_1 = 2$, $R_2 = 5$ and $R_3 = 3$. The *Mann-Whitney test* can then be determined by defining a rejection region based on the statistic $T = \sum_{i=1}^{n} R_i$ [NOTE: this test is also sometimes referred to as the *Wilcoxon rank-sum test*]. It is a reasonably straightforward (though tedious) exercise to show that, under the null hypothesis, the mean and variance of T are:

$$E_{H_0}(T) = \frac{n(n+m+1)}{2}; \qquad Var_{H_0}(T) = \frac{nm(n+m+1)}{12}.$$

Now, if the observed value of T is far from its expectation under the null hypothesis, then this is evidence that we should reject the null hypothesis. Indeed, the rejection region for the Mann-Whitney test is of the form $C = \{|T - E_{H_0}(T)| \geq k\}$. All that remains is to appropriately determine the value k to ensure the desired size of the test. For the simple data set with $n = 3$ and $m = 2$ given earlier, we note that the observed value of T is $2 + 5 + 3 = 10$. To determine the distribution of T in this case, we note that for $n = 3$ and $m = 2$, there are 10 possible general arrangements for the combined values in terms of the sample to which the values belong; that is, the ordered sample could have been associated with the arrangements:

$$xxxyy,\ xxyxy,\ xxyyx,\ xyxxy,\ xyxyx,\ xyyxx,\ yxxxy,\ yxxyx,\ yxyxx,\ yyxxx$$

(e.g., the given data are in the arrangement $yxxyx$). For each of these 10 arrangements, the associated values of T are 6, 7, 8, 8, 9, 10, 9, 10, 11, and 12. Under the null hypothesis, each of these 10 arrangements is equally likely, and thus we can calculate:

$$Pr_{H_0}(T \leq 6) = \frac{1}{10}, \qquad Pr_{H_0}(T \leq 7) = \frac{1}{5}, \qquad Pr_{H_0}(T \geq 11) = \frac{1}{5}, \qquad Pr_{H_0}(T \geq 12) = \frac{1}{10}.$$

As such, if we want a test with size $\alpha = 0.2$, we could use the rejection region $C = \{T = 6 \textit{ or } T = 12\}$. [NOTE: Again, we see that it is not always possible to construct tests for all possible sizes.] Unfortunately, the exact distribution of the statistic T is quite complicated when m and n are reasonably large. However, as for the Wilcoxon signed-rank statistic, it turns out that, under the null hypothesis, the distribution of T is well-approximated by a normal distribution with mean $E_{H_0}(T)$ and variance $Var_{H_0}(T)$. As such, we can define a rejection region for the Mann-Whitney test with approximate size α by setting $k = \Phi^{-1}(1 - \alpha/2)\sqrt{Var_{H_0}(T)}$.

Example 4.6: Suppose that we observe two samples of size $n = 10$ and $m = 9$ as follows:

$$X: \ 4.3,\ 5.9,\ 4.9,\ 3.1,\ 5.3,\ 6.4,\ 6.2,\ 3.8,\ 7.1,\ 5.8,$$
$$Y: \ 5.5,\ 7.9,\ 6.8,\ 9.0,\ 5.6,\ 6.3,\ 8.5,\ 4.6,\ 7.5.$$

The sorted combined sample (along with whether the observations was an X_i or a Y_j) is:

$$3.1(x),\ 3.8(x),\ 4.3(x),\ 4.6(y),\ 4.9(x),\ 5.3(x),\ 5.5(y),\ 5.6(y),\ 5.8(x),\ 5.9(x),$$
$$6.2(x),\ 6.3(y),\ 6.4(x),\ 6.8(y),\ 7.1(x),\ 7.5(y),\ 7.9(y), 8.5(y),\ 9.0(y).$$

So, the ranks of the X_i's are seen to be $R_1 = 3$, $R_2 = 10$, $R_3 = 5$, $R_4 = 1$, $R_5 = 6$, $R_6 = 13$, $R_7 = 11$, $R_8 = 2$, $R_9 = 15$, $R_{10} = 9$, implying that the observed value of the test statistic is

$T = 75$. Furthermore, we see that under the null hypothesis, the mean and variance of T are given by

$$E_{H_0}(T) = \frac{10(10+9+1)}{2} = 100; \qquad Var_{H_0}(T) = \frac{10(9)(10+9+1)}{12} = 150.$$

Therefore, a size $\alpha = 0.05$ test is determined by the critical region $C = \{|T - 100| \geq k\}$, where $k = 1.96\sqrt{150} = 24.005$. So, since $|T - 100| = |75 - 100| = 25$, we reject the null hypothesis (of course, if we had desired a test of size $\alpha = 0.01$ we would not have rejected H_0, since in this case the appropriate value of k would have been $2.575\sqrt{150} = 31.537$). Finally, by way of comparison, we note that the observed number of X_i's less than the combined sample median of 5.9 is $V = 6$. Using the normal approximation to the distribution of V, we see that a median test with size $\alpha = 0.05$ is determined by the critical region $C = \{|V - 5| \geq 1.96\sqrt{1.25}\} = \{|V - 5| \geq 2.19\}$, since

$$E_{H_0}(V) = \frac{10}{2} = 5; \qquad Var_{H_0}(V) = \frac{9(10)}{4(9+10-1)} = 1.25.$$

Thus, since $|V - 5| = 1$ in this case, we do not reject the null hypothesis. This is a nice example of how the median test is less powerful than the Mann-Whitney test (not an overly surprising result given that the median test ignores more information contained in the observed data than does the Mann-Whitney test).

As noted at the end of Example 4.6, the Mann-Whitney test is generally more powerful than the median test since it takes into account, to some degree, the relative sizes of the observed data values. Of course, it only takes account of these sizes through the use of ranks, and thus still ignores some potentially relevant information. As such, we close this section by introducing another testing procedure for the null hypothesis $H_0 : F(z) = G(z)$ for all z which does take into account the actual observed values of the data directly.

Supposing that we have observed two independent samples, $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ and we have settled on some statistic $T = T(X, Y)$ which can be used to investigate the potential differences between the two samples (e.g., the most common choice would be $\overline{X} - \overline{Y}$, though many other choices are possible). As in the parametric setting, we then construct a test based on a rejection region of the form $C = \{T \leq k_1 \text{ or } T \geq k_2\}$ for values of k_1 and k_2 chosen to ensure that the size of the resulting test was some desired value α. In the parametric setting, we would use our chosen underlying probability model to determine the value of k_1 and k_2. However, in the current setting, we have avoided making parametric assumptions. Nonetheless, it is possible to determine values of k_1 and k_2 under the assumption of the null hypothesis of equal distributions within the two populations under study. We note that if H_0 is true, then the observation labels (i.e., whether the observation is associated with the X-sample or the Y-sample) are equally likely to have arisen in any of the $\binom{n+m}{n}$ possible allocations of the observed values to X and Y samples. As such, we can define a new data set $X^\star = (X_1^\star, \ldots, X_n^\star)$, $Y^\star = (Y_1^\star, \ldots, Y_m^\star)$, which is just a permutation of the original samples (so that values in the original X-sample may now appear in the new Y-sample instead) and a new test statistic value $T^\star = T(X^\star, Y^\star)$. If we calculate values of $T^\star$ for all of the possible re-allocations of the data labels, then we can approximate the probability $Pr_{H_0}(C)$ by simply calculating the proportion of these $T^\star$ values which fall in the set C. Or, conversely, we can construct a rejection region with (approximately) the desired size α by selecting k_1 and k_2 to be the lower and upper $\alpha/2$-quantiles of the observed distribution of the $T^\star$ values; that is, if we represent the ordered collection of the $N = \binom{n+m}{n}$ $T^\star$ values as $T^\star_{[1]}, \ldots, T^\star_{[N]}$, then $k_1 = T^\star_{[N\alpha/2]}$ and $k_2 = T^\star_{[N(1-\alpha/2)]}$ [where, of course, we must round off the values $N\alpha/2$ and

$N(1-\alpha/2)$ to the nearest integer value]. The test so constructed is often referred to as a *permutation test*, due to the process of permutating of sample labels on which it is based. We stress that, despite the fact that we are using the actual observed values of our data in the construction of the test, their are no parametric assumptions being employed. The actual implementation of the testing process is easiest to understand by examination of a simple example:

Example 4.6: Suppose that we observed the two datasets

$$X_1 = 4, X_2 = 3, X_3 = 7; \qquad Y_1 = 1, Y_2 = 9.$$

Further, suppose that we use the standard statistic $T = \overline{X} - \overline{Y}$ to distinguish between the two samples. We note that there are $\binom{5}{3} = 10$ different re-allocations of the data values into an X-sample of size 3 and a Y-sample of size 2. These 10 re-allocations, along with their associated value of $T^\star$ are:

$$\begin{array}{ll}
X_1^\star = 1, X_2^\star = 3, X_3^\star = 4, Y_1^\star = 7, Y_2^\star = 9: & T^\star = 2.67 - 8.0 = -5.33 \\
X_1^\star = 1, X_2^\star = 3, X_3^\star = 7, Y_1^\star = 4, Y_2^\star = 9: & T^\star = 3.67 - 6.5 = -2.83 \\
X_1^\star = 1, X_2^\star = 3, X_3^\star = 9, Y_1^\star = 4, Y_2^\star = 7: & T^\star = 4.33 - 5.5 = -1.17 \\
X_1^\star = 1, X_2^\star = 4, X_3^\star = 7, Y_1^\star = 3, Y_2^\star = 9: & T^\star = 4.00 - 6.0 = -2.00 \\
X_1^\star = 1, X_2^\star = 4, X_3^\star = 9, Y_1^\star = 3, Y_2^\star = 7: & T^\star = 4.67 - 5.0 = -0.33 \\
X_1^\star = 1, X_2^\star = 7, X_3^\star = 9, Y_1^\star = 3, Y_2^\star = 4: & T^\star = 5.67 - 3.5 = 2.17 \\
X_1^\star = 3, X_2^\star = 4, X_3^\star = 7, Y_1^\star = 1, Y_2^\star = 9: & T^\star = 4.67 - 5.0 = -0.33 \\
X_1^\star = 3, X_2^\star = 4, X_3^\star = 9, Y_1^\star = 1, Y_2^\star = 7: & T^\star = 5.33 - 4.0 = 1.33 \\
X_1^\star = 3, X_2^\star = 7, X_3^\star = 9, Y_1^\star = 1, Y_2^\star = 4: & T^\star = 6.33 - 2.5 = 3.83 \\
X_4^\star = 1, X_2^\star = 7, X_3^\star = 9, Y_1^\star = 1, Y_2^\star = 3: & T^\star = 6.67 - 2.0 = 4.67.
\end{array}$$

Since each of these $T^\star$ values is equally likely under the null hypothesis, we see that the region $C = \{T \le -5.33 \textit{ or } T \ge 4.67\}$ has an approximate size of 0.2 (since 2 of the ten re-allocations yield $T^\star$ values which lie in C). As such, we have constructed a test with size $\alpha = 0.2$. Since our observed value is $T = -0.33 \notin C$, we see that we cannot reject the null hypothesis $H_0 : F(z) = G(z)$ for all z. Of course, we could just as easily used some other statistic T', say the difference in medians. In general, the choice of statistic will depend upon how we believe the two populations are likely to differ from one another, and thus is a quite problem specific issue.

In general, when $n + m$ is large, the number of re-allocations of the labels is extremely large (e.g., for the dataset of Example 4.5, where $n = 10$ and $m = 9$, there are 92,378 different re-allocations of the data into two samples of appropriate size). In such cases, it is common practice to use only a random subset of some number B of the possible re-allocations. We note the similarity in this regard to the idea underlying the bootstrap introduced in Section 2.6.3. Indeed, the bootstrap can also be used to construct non-parametric hypothesis tests, but we do not discuss this idea here.