Hypothesis Testing

1 Introduction

This document is a simple tutorial on hypothesis testing. It presents the basic concepts and definitions as well as some frequently asked questions associated with hypothesis testing. Most of the material presented has been taken directly from either Chapter 4 of Scharf [3] or Chapter 10 of Wasserman [4].

2 Hypotheses

Let $\mathbf{X}$ be a random vector with the range χ and distribution $F_\theta(\mathbf{x})$. The parameter θ is belongs to the parameter space Θ. Let $\Theta = \Theta_0 \bigcup \Theta_1 \bigcup ... \bigcup \Theta_{M-1}$ be a disjoint covering of the parameter space. We define H_i as the hypothesis is that $\theta \in \Theta_i$.

An **M-ary hypothesis test** chooses which of the M disjoint subsets contain the unknown parameter θ. When $M = 2$ we have a **binary hypothesis test**. For the remainder of this document we will only discuss binary hypothesis tests ($H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$).

2.1 Null and Alternative Hypotheses

In many binary hypothesis tests H_0 is referred to as the **null hypothesis** and H_1 the **alternative hypothesis**. This is due to the fact that in most binary tests the hypothesis H_0 is set up to be refuted in order to support an alternative hypothesis H_1. That is, H_0 usually represents the absence of some effect/factor/condtion. For example when testing if a new drug is better than a placebo for relieving a set of symptoms the null hypothesis H_0 says the new drug has the same effect as the placebo. With tests of this form it is common to talk about a hypothesis test in terms accepting or rejecting the null hypothesis.

2.2 Simple vs. Composite

When Θ_i contains a single element θ_i hypothesis H_i is said to be **simple**. Otherwise it is **composite**. A binary hypothesis test can be simple vs simple, simple vs composite, composite vs simple, or composite vs composite. Here are some simple examples:

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta = \theta_1 \quad \text{(simple vs simple)}$$

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta \neq 0 \quad \text{(simple vs composite)}$$

$$H_0 : \theta < 0 \text{ versus } H_1 : \theta > 0 \quad \text{(composite vs composite)}$$

2.3 One-Sided and Two-Sided Tests

A hypothesis test is considered to be **two-sided** if the is of the form:

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

where the alternative hypothesis H_1 "lies on both sides of H_0". A test of the form

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0$$

is called a **one-sided** test. Note that one-sided and two-sided tests are only defined for scalar parameter spaces and at least one hypothesis must be composite.

2.4 Frequentist vs. Bayesian View of Hypotheses

Thus far we discussed hypothesis testing in terms of determining which subset of a parameter space an unknown θ lies. The classic/frequentist approach to hypothesis testing treats θ as deterministic but unknown. A Bayesian approach treats θ as a random variable and assumes there is a distribution on the possible θ in the parameter space. That is, one can define a prior on each hypothesis being true. Discussion of advantageous and disadvantageous of each of these will be spread throughout the following sections.

3 Binary Hypothesis Testing

Give a sample $\mathbf{x}$ of a random vector $\mathbf{X}$ whose range is χ and has the distribution $F_\theta(\mathbf{x})$ a binary hypothesis test ($H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$) takes the form

$$\phi(\mathbf{x}) = \begin{cases} 0 \sim H_0, & \mathbf{x} \in A \\ 1 \sim H_1, & \mathbf{x} \in A^c, \end{cases} \tag{1}$$

This equation is read as, "the test function $\phi(\mathbf{x})$ equals 0, and the hypothesis H_0 is accepted, if the measurement $\mathbf{x}$ lies in the acceptance region A (where $A \subset \chi$). If the measurement lies outside this region then test function equals 1 and hypothesis H_0 is rejected and H_1 is accepted." Usually the region A is of the form:

$$A = \{\mathbf{x} : T(\mathbf{x}) < c\} \tag{2}$$

where T is a **test statistic** and c is a critical value. The trick is to find the appropriate test statistic T and an appropriate critical value c. We will be more explicit about what "appropriate" means in the following sections.

3.1 Type I and Type II Errors

There are two types of errors a binary hypothesis test can make. A **type I error** or **false alarm** is when H_0 is true, but $\mathbf{x} \in A^c$. That is, the test chooses H_1 when H_0 is true. A **type II error** or **miss** is when H_1 is true, but $\mathbf{x} \in A$. That is, the test chooses H_0 when H_1 is true.

3.2 Size and Power

If H_0 is simple ($\Theta_0 = \{\theta_0\}$), the **size** or probability of false alarm is

$$\alpha = P_{\theta_0}(\phi(\mathbf{x}) = 1) = E_{\theta_0}[\phi(\mathbf{x})] = P_{FA}. \tag{3}$$

where $E_{\theta_0}[\phi(\mathbf{x})]$ indicates that $\phi(\mathbf{x})$ is averaged under the density function $f_{\theta_0}(\mathbf{x})$.

If H_0 is composite, the size is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} E_\theta[\phi(\mathbf{x})]. \tag{4}$$

The size is the worst-case probability of choosing H_1 when H_0 is true. A test is said to have **level** α if its size is less than or equal to α.

A hit or detection is when H_1 is true, and $\mathbf{x} \in A^c$. If H_1 is simple, the **power** or probability of detection is

$$\beta = P_{\theta_1}(\phi(\mathbf{x}) = 1) = E_{\theta_1}[\phi(\mathbf{x})] = P_D. \tag{5}$$

If H_1 is composite, the power is defined for each $\theta \in \Theta_1$ as $\beta(\theta)$. In fact everything can be defined in terms of this power function:

$$\beta(\theta) = P_\theta(\phi(\mathbf{x}) = 1) = P_\theta(\mathbf{x} \in A^c) \tag{6}$$

That is the power of a composite test is defined for each $\theta \in \Theta_1$ and the size can be written as:

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) \tag{7}$$

A receiver operating characteristic (ROC) curve is a plot of β versus α (for a simple vs simple hypothesis test). Usually multiple β, α pairs are obtained by adjusting the threshold / critical value c in Equation 2.

3.3 Bias

A test $\phi(\mathbf{x})$ is said to be unbiased if its power is never smaller than its size. That is,

$$\beta(\theta) \geq \alpha \; \forall \; \theta \in \Theta_1 \tag{8}$$

3.4 Best and Uniformly Most Powerful Test

For a simple versus simple binary hypothesis test, $\phi(\mathbf{x})$ is the best test of size α if it has the most power among all tests of size α. That is, if $\phi(\mathbf{x})$ and $\phi'(\mathbf{x})$ are two competing tests each of which has size α, then $\beta \geq \beta'$. The **best** test maximizes the probability of detection (power) for a fixed probability of false alarm (size). Neyman-Pearson will show us the form of the best test for a fixed α in the next section.

A test $\phi(\mathbf{x})$ is said to be **uniformly most powerful (UMP)** of size α if it has size α and its power is uniformly (for all θ) greater than the power of any other test $\phi'(\mathbf{x})$ whose size is less than or equal to α. That is:

$$\sup_{\theta \in \Theta_0} E_\theta[\phi(\mathbf{x})] = \alpha$$

$$\sup_{\theta \in \Theta_0} E_\theta[\phi'(\mathbf{x})] \leq \alpha$$

$$E_\theta[\phi(\mathbf{x})] \geq E_\theta[\phi'(\mathbf{x})] \; \forall \; \theta \in \Theta_1$$

In general a UMP test may be difficult to find or may not exist. One strategy to proving a test is UMP is to find the best test for a particular θ and then show the test does not depend on θ. The Karlin-Rubin theorem shows how to obtain the UMP test for certain one-sided hypothesis tests (See [3]).

4 Neyman-Pearson Lemma

The Neyman-Pearson Lemma shows how to find the most powerful or best test of size α when H_0 and H_1 are both simple. The lemma tells us the "appropriate" test statistic T to maximize the power given a fixed α. The test is a slight generalization of the test defined in 1. The lemma states that:

$$\phi(\mathbf{x}) = \begin{cases} 1, f_{\theta_1}(\mathbf{x}) > k f_{\theta_0}(\mathbf{x}) \\ \gamma, f_{\theta_1}(\mathbf{x}) = k f_{\theta_0}(\mathbf{x}) \\ 0, f_{\theta_1}(\mathbf{x}) < k f_{\theta_0}(\mathbf{x}), \end{cases} \tag{9}$$

or alternatively

$$\phi(\mathbf{x}) = \begin{cases} 1, T(\mathbf{x}) > k \\ \gamma, T(\mathbf{x}) = k \\ 0, T(\mathbf{x}) < k \end{cases} \tag{10}$$

for some $k \geq 0$, $0 \leq \gamma \leq 1$, is the most powerful test of size $\alpha > 0$ for testing H_0 versus H_1. $T(\mathbf{x}) = f_{\theta_1}(\mathbf{x}) / f_{\theta_0}(\mathbf{x}) = L(\mathbf{x})$ and is called the **likelihood ratio**. When $\phi(\mathbf{x})$ is 1 or 0 it is the same as in Equation 1. However, when $\phi(\mathbf{x}) = \gamma$ we "flip a γ coin" to select H_1 with probability γ (when the coin comes up heads).

Proof. Consider any test $\phi'(\mathbf{x})$ such that $\alpha' \leq \alpha$. We have

$$\int [\phi(\mathbf{x}) - \phi'(\mathbf{x})][f_{\theta_1}(\mathbf{x}) - k f_{\theta_0}(\mathbf{x})] \geq 0 \tag{11}$$

$$\beta - \beta' \geq k(\alpha - \alpha') \tag{12}$$

$$\geq 0. \tag{13}$$

□

4.1 Setting the size to α

The question remains of how to set k to produce a test of size α. The size for this test is:

$$\alpha = E_{\theta_0}[\phi(\mathbf{x})] = P_{\theta_0}[L(\mathbf{x}) > k] + \gamma P_{\theta_0}[L(\mathbf{x}) = k] \tag{14}$$

If there exists a k_0 such that

$$P_{\theta_0}[L(\mathbf{x}) > k_0] = \alpha \tag{15}$$

then we set $\gamma = 0$ and pick $k = k_0$. Otherwise there exists a k_0' such that

$$P_{\theta_0}[L(\mathbf{x}) > k_0'] \leq \alpha < P_{\theta_0}[L(\mathbf{x}) \geq k_0'] \tag{16}$$

We can then use $k = k_0'$ and choose the γ that solves:

$$\gamma P_{\theta_0}[L(\mathbf{x}) = k_0'] = \alpha - P_{\theta_0}[L(\mathbf{x}) > k_0'] \tag{17}$$

4.2 Interpretation

So the Neyman-Pearson tells us that the best / most powerful test for a fixed size alpha is one that uses that makes decisions by thesholding the likelihood ratio $L(\mathbf{x})$. We refer to such tests as **likelihood ration tests (LRT)** . Note that the test statistic is the likelihood ratio $L(\mathbf{x})$ and is a random variable. If $L(\mathbf{x}) = k$ with probability zero (which is most likely the case for continuous $\mathbf{x}$) then the threshold k is found as:

$$\alpha = P_{\theta_0}[L(\mathbf{x})] > k] = \int_k^\infty f_{\theta_0}(L) dL \tag{18}$$

where $f_{\theta_0}(L)$ is the density function for $L(\mathbf{x})$ under H_0.

5 Bayesian Hypothesis Testing

In the previous sections we discussed simple binary hypothesis testing in the following framework. Given a measurement $\mathbf{x}$ drawn from the distribution $F_\theta(\mathbf{x})$, how do we choose whether $\theta = \theta_0$ or $\theta = \theta_1$. We defined hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ and look for a test of H_1 versus H_0 that is "optimal". We talked about optimality in terms of maximizing the power (β) of such a test for a fixed size α. The parameter θ (and the hypothesi) are treated at deterministic but unknown quantities. That is either H_1 or H_0 is true and we don't know which one. We don't have any prior knowledge of how likely H_1 or H_0 is to occur or how likely any parameter choice is. Note that the power and size are both defined in terms of one of the hypothesis being true.

The Bayesian approach to hypothesis testing treats θ and the hypothesis H as unknown random variables. Here we are introducing the random variable H to represent the hypothesis. If $H = i$ it means hypothesis H_i is true. We can think of the test function ϕ as an estimator for H. The conceptual framework is as follows. Mother Nature selects H and the parameter θ from a joint distribution $f(\theta, H) = p(H)f(\theta|H)$. She does this by first choosing H according to $p(H)$ and then picks a θ according to $f(\theta|H)$. Note that $f(\theta|H = i) = 0$ for all $\theta \notin \Theta_i$. Her selection determines from which distribution $F_\theta(\mathbf{x})$ Father Nature draws his measurement. This measurement is given to the experimenter and he or she must decide between estimate the value of H via a decision function ϕ. Each time the experiment is run a parameter θ is chosen by Mother nature, and the

experimenter outputs a decision $\hat{H} = \phi(\mathbf{x})$. The goal in Bayesian hypothesis testing to design a test ϕ that is "optimal" / gives the best performance. Here "optimality" is described in terms of the Bayes risk which is described below.

To be more concrete let us consider the simple versus simple binary hypothesis test we have been discussing so far. Let $p(H = H_0) = p_0$ and $p(H = H_1) = 1 - p_0$. Since the hypothesis is simple $f(\theta|H = H_i) = \delta(\theta - \theta_i)$. Mother nature selects H according to $p(H)$. In the simple binary case we are considering this is equivalent to picking $\theta = \theta_0$ with probability p_0 and $\theta = \theta_1$ with probability $p_1 = 1 - p_0$. Depending on her choice Father Nature then draws a a measurement $\mathbf{x}$ from either $F_{\theta_0}(\mathbf{x})$ or $F_{\theta_1}(\mathbf{x})$. Our goal will be to obtain $\hat{H} = \phi(\mathbf{x})$ that minimizes the Bayes risk.

5.1 Cost of Decisions

A cost or loss function is defined for each possible pairing of the true hypothesis H and decision $\hat{H} = \phi(\mathbf{x})$. That is for the pair $(H, \phi(\mathbf{x}))$ = (i, j) we assign a nonnegative cost $C[H = i, \phi(\mathbf{x}) = j] = c_{ij}$. We say c_{ij} the loss incurred when Mother Nature selects the hypothesis H_i and the experimenter decides to choose H_j. That is for a simple binary hypothesis test we give values for c_{00}, c_{11}, c_{01} and c_{10}. Normal we do not associate a loss/cost for making a correct decision, *i.e.* $c_{00} = c_{11} = 0$.

5.2 Risk

We define the risk $R(H_i, \phi(\mathbf{x}))$ for each H_i as the expected loss given $H = H_i$ for particular test function ϕ:

$$R(H, \phi) = E_{\mathbf{x}}[C[H, \phi(\mathbf{x})]] = \begin{cases} c_{00}P_{00} + c_{01}P_{01}, \ \theta = \theta_0 \\ c_{10}P_{10} + c_{11}P_{11}, \ \theta = \theta_1 \end{cases} \tag{19}$$

where $P_{ij} = p(\phi(\mathbf{x}) = 1|H = H_i)$. This is equivalent to $p_{\theta_i}(\phi(\mathbf{x}) = j)$ in the simple binary hypothesis case.

5.3 Bayes Risk

The Bayes risk is the average risk over the distribution of H that Mother Nature used (for the binary case, only $P(H = 0) = p_0$ is needed).

$$R(p_0, \phi) = E_H[R(H, \phi)] = p_0 R(H = 0, \phi) + (1 - p_0)R(H = 1, \phi) \tag{20}$$

Given that we known the prior p_0, the optimal test ϕ is defined to be the one that minimizes the Bayes Risk:

$$\phi = \arg\min_{\phi'} R(p_0, \phi') \tag{21}$$

It turns out that the solution to this equation/optimization has the form (see 6.432 notes or Scharf [3] Chapter 5):

$$\phi(\mathbf{x}) = \begin{cases} 1, L(\mathbf{x}) > \eta \\ 0, L(\mathbf{x}) < \eta \end{cases} \tag{22}$$

where

$$L(\mathbf{x}) = \frac{f(\mathbf{x}|H = 1)}{f(\mathbf{x}|H = 0)} = \frac{\int_\theta f(\mathbf{x}|H = 1, \theta) f(\theta|H = 1) d\theta}{\int_\theta f(\mathbf{x}|H = 0, \theta) f(\theta|H = 0) d\theta} \tag{23}$$

is the likelihood ratio and for the simple versus simple case is

$$L(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} = \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} \tag{24}$$

and the threshold

$$\eta = \frac{p_0(c_{10} - c_{00})}{(1 - p_0)(c_{01} - c_{11})} \tag{25}$$

So once again we see that the "optimal" test is a likelihood ratio test. Here "optimal" is in terms of minimizing Bayes risk. The test statistic is the likelihood ratio and the acceptance region depends on the threshold η which is based on the priors and decision costs.

Minimizing risk sounds like the right way to think about these problems. However this approach requires us to have some knowledge about the prior on θ and be able to assign a cost to each possible outcome. This may be difficult in some applications. For example, what is the prior probability of a missile coming toward your home? We will quickly discuss what can be done when we can assign costs to decision but don't know the a priori probabilities of each hypothesis in the Minmax section below. When we don't know the prior AND cannot think of a meaningful costs we go back to Neyman-Pearson testing.

6 Test Statistics and Sufficiency

When we talk about the hypothesis H being a random variable we can consider the following Markov chain $H \rightarrow \mathbf{X} \rightarrow T(\mathbf{X})$ where $T(\mathbf{X})$ is some test statistic. Any test statistic is said to be sufficient for H if $p(H|\mathbf{X}) = p(H|T(\mathbf{X}))$ That is, if $T(\mathbf{X})$ is sufficient it tells us everything we need to know about the observation $\mathbf{x}$ in order to estimate H and the Markov chain can be written as $H \rightarrow T(\mathbf{X}) \rightarrow \mathbf{X}$.

Note that the likelihood ratio $L(\mathbf{X})$ was to optimal test statistic for our hypothesis test. It takes our K dimension observation $\mathbf{x}$ and maps it to a scalar value. It can be shown that $L(\mathbf{X})$ is a sufficient statistic for H.

This was just a quick note. More details can be found in the 6.432 notes.

7 Minimax Tests

As we eluded to before we may be able to associated a cost with each of the possible outcomes of a hypothesis test but have no idea what the prior on H is (or even worse the full $f(\theta, H)$). In such cases we can play a game in which we assume Mother Nature is really mean and will choose a prior that makes whatever test we choose to look bad. That is for a simple versus simple binary hypothesis test:

$$max_{p_o} min_{\phi} R(p_0, \phi) \tag{26}$$

To combat this, we will try to find a ϕ that minimizes the worst she can do:

$$min_{\phi} max_{p_o} R(p_0, \phi) \tag{27}$$

Seciont 5.3 of Scharf [3] shows how to find such a minmax detector / test function ϕ. It is also shown that

$$max_{p_o} min_{\phi} R(p_0, \phi) = min_{\phi} max_{p_o} R(p_0, \phi) \tag{28}$$

This topic is also discussed in the 6.432 notes.

8 Generalized Likelihood Ratio Test

A discussed before when dealing with composite hypotheses in Neyman-Pearson framework we wish to find the UMP test for a fixed size α (put your Bayes hat away for a bit). However, it is typically the case that such a test does not exist. A **generalized likelihood ratio test** is a way to deal with general composite hypothesis test. Again we will focus on the binary case with $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The generalized likelihood ratio is a test statistic with the following form:

$$L_G(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_1} f(\mathbf{x}|\theta)}{\sup_{\theta \in \Theta_0} f(\mathbf{x}|\theta)} \tag{29}$$

We see that for a simple versus simple hypothesis test $L_G(\mathbf{x}) = L(\mathbf{x})$. This test statistic is rather intuitive. If top part of the fraction in Equation 29 is greater than the bottom part then the data is best explained when

$\theta \in \Theta_1$. If the opposite is true then the data is best explained by $\theta \in \Theta_0$. Instead of using Equation 29 as the test statistic one generally uses:

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta} f(\mathbf{x}|\theta)}{\sup_{\theta \in \Theta_0} f(\mathbf{x}|\theta)} \tag{30}$$

instead, where $\Theta = \Theta_0 \bigcup \Theta_1$. It can be shown in general that $\lambda(\mathbf{x}) = \max(L_G(\mathbf{x}), 1)$. This may seem a little strange in that Θ and Θ_0 are nested (they are not disjoint). Much more can be said about this, but for now I will simply paraphrase Wasserman's tutorial on this subject [4] and say that in practice using $\lambda(\mathbf{x})$ instead of $L_G(\mathbf{X})$ has little effect in practice and theoretically properties of $\lambda(\mathbf{x})$ are usually much easier to obtain.

So the generalized likelihood ratio test gives us a test statistic that makes some intuitive sense. We can threshold this statistic and calculate it's size and power if we can derive it's distribution under each hypothesis.

Now let's put our Bayes hat back on. We showed before that the likelihood ratio test minimizes the Bayes risk and that for a composite test the likelihood ratio is:

$$L(\mathbf{x}) = \frac{\int_\theta f(\mathbf{x}|H=1,\theta) f(\theta|H=1) d\theta}{\int_\theta f(\mathbf{x}|H=0,\theta) f(\theta|H=0) d\theta} \tag{31}$$

That is we should integrate over all possible values of theta for each hypothesis in each hypothesis. However, typically our parameter space is extremely large. If we assume that $f(\mathbf{x}|H=1,\theta)f(\theta|H=1)$ contains a large peak (looks like a delta function) at $\hat{\theta_1} = \arg\max_{\theta \in \Theta_1} f(\mathbf{x}|H=1,\theta)f(\theta|H=1)$ and $f(\mathbf{x}|H=1,\theta)f(\theta|H=0)$ peaks at $\hat{\theta_0}$ we can approximate the likelihood ratio as:

$$\hat{L}(\mathbf{x}) = \frac{f(\mathbf{x}|\hat{\theta_1}) f(\hat{\theta_1}|H=1)}{f(\mathbf{x}|\hat{\theta_0}) f(\hat{\theta_1}|H=0)} = \frac{\max_{\theta \in \Theta_1} f(\mathbf{x}|H=1,\theta) f(\theta|H=1)}{\max_{\theta \in \Theta_0} f(\mathbf{x}|H=0,\theta) f(\theta|H=0)} \tag{32}$$

which is one possible interpretation of the generalized likelihood ratio $\hat{L}_G(\mathbf{x})$ (ignore the details involved with $\max$ and $\sup$).

9 Test of Significance

In the Neyman-Pearson testing framework one fixes the size of the α of the test. However, different people may have different criteria for choosing an appropriate size. One experimenter may be happy with setting the size to $\alpha = 0.05$ while another demands α is set to 0.01. In such cases it is possible that one experimenter accepts H_0 while the other rejects it when given the same data $\mathbf{x}$. Only reporting if H_0 was accepted or rejected may not be very informative in such cases. If the experimenters both use the same test statistic (*i.e.* both do a likelihood ratio test) it may be more useful for them to report the outcome of their experiment in terms of the **significance** probability or **p-value** of the test (also referred to as the **observed size**). We give a formal definition for the p-value below.

9.1 P-value

If a test rejects H_0 at a level α it will also reject at a level $\alpha' > \alpha$. Remember that when a test rejects H_0 at level α that means it's size is less than or equal to α. (*i.e.* if we say test rejects at level .05 that means it's size $\alpha \leq .05$). The **p-value** is the smallest α at which a test rejects H_0.

Suppose that for every $\alpha \in (0,1)$ we have a size α test with a rejection A_α^c, then

$$\text{p-value} = \inf\{\alpha : T(\mathbf{x}) \in A_\alpha^c\} \tag{33}$$

That is the p-value is the smallest level α at which an experimenter using the test statistic T would reject H_0 on the basis of the observation $\mathbf{x}$.

Ok, that definition requires a lot of thought to work through. It may be easier to understand what a p-value is if we explain how to calculate one:

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T(\mathbf{X}) \geq T(\mathbf{x})) \tag{34}$$

where $\mathbf{x}$ is the observed value of $\mathbf{X}$. If H_0 is a simple hypothesis and $\Theta_0 = \{\theta_0\}$ then

$$\text{p-value} = P_\theta(T(\mathbf{X}) \geq T(\mathbf{x})) = P(T(\mathbf{X}) \geq T(\mathbf{x})|H_0) \tag{35}$$

It is important remember that $T(\mathbf{X})$ is a random variable with a particular distribution under H_0 and that $T(\mathbf{x})$ is a number, the value of the test statistic for the observed $\mathbf{x}$. In the case of a simple H_0 the p-value is the probability of obtaining a test statistic value greater than the one you observed when H_0 is true. Another way to look at it is that the p-value is the size of a test using your observed $T(\mathbf{x})$ as the threshold for rejecting H_0.

If the test statistic $T(\mathbf{X})$ has a continuous distributen then under a simple $H_0 : \theta = \theta_0$, the p-value has a uniform distribution between 0 and 1. If we reject H_0 when the p-value is less than α then the probability of false alarm (or size of the test) is α. That is we can set up a test to have size α by making the test function be:

$$\phi(\mathbf{x}) = \begin{cases} 1, \text{p-value} \geq \alpha \\ 0, \text{p-value} < \alpha, \end{cases} \tag{36}$$

So a small p-value is strong evidence against H_0. However, note that **a large p-value is NOT strong evidence in favor of** H_0. A large p-value can mean H_0 is true OR H_0 is false but the test has low power β. It also important to note that the **p-value is not the probability that the null hypothesis is true** . That is, in almost every case the p-value $\neq p(H_0|\mathbf{x})$. We will shown one exception in the next section.

9.2 P-values for One-Sided Tests

Let's look at a one-sided tests where $\Theta \subset R$ and θ_0 is entirely to one side of θ_1. In this case, p-values will sometimes have a Bayesian justification. For example if $\mathbf{X} \sim N(\theta, \sigma^2)$ and $p(\theta) = 1$, then $p(\theta|\mathbf{x})$ is $N(\mathbf{x}, \sigma^2)$. We test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$,

$$p(H_0|\mathbf{x}) = p(\theta \leq \theta_0|\mathbf{x}) = \Phi\left(\frac{\theta_0 - \mathbf{x}}{\sigma}\right). \tag{37}$$

The p-value is

$$\text{p-value} = p(\mathbf{X} \geq \mathbf{x}) = 1 - \Phi\left(\frac{\mathbf{x} - \theta_0}{\sigma}\right) = p(H_0|\mathbf{x}) \tag{38}$$

because Φ is symmetric.

10 Permutation Tests

We showed how to calculate a p-value or significance in the previous section. This calculation requires knowing the distribution (the cdf) of the test statistic $T(\mathbf{X})$ under H_0. However, in many cases it may be difficult to obtain this distribution (*i.e.* distribution of $\mathbf{X}$ may be unknown). A **permutation tests** are tests based on non-parametric estimates of significance. It does not rely on the distribution of the test statistic. The basic procedure is as follows:

1. Compute the observed value of the test statistic $t_{obs} = T(\mathbf{x})$
2. Obtain a new sample $\mathbf{x}_s$ that *obeys* the null hypothesis H_0 via a resampling function π.
   That is $\mathbf{x}_s = \pi(\mathbf{x})$.
3. Compute $t_s = T(\mathbf{x}_s)$
4. Repeat Steps 2 and 3 B times and let $t_1, ..., t_B$ denote the resulting values.
5. Calculate an approximate $\hat{p}$-value $= \frac{1}{B}\sum_{j=1}^{B} I(t_j > t_{obs})$ where $I(true) = 1$ and $I(false) = 0$.
6. Reject H_0 (choose H_1) if $\hat{p}$-value $> \alpha$

where step 5 uses the empirical cumulative distribution obtained from samples in step 2 to estimate the p-value. The question remains on what we mean by *obeys* the null hypothesis in step 2. The resampling function π obeys H_0 if each new sample $\mathbf{x}_s$ is equality likely when H_0 is true.

Take for example an observation $\mathbf{x} = (y_1, y_2, ..., y_N, z_1, ..., z_M)$ which is N observations of some random variable $\mathbf{Y}$ and M observations of the random variable $\mathbf{Z}$. If H_0 was that both Z and Y have the same mean one possible test statistic would be $T(\mathbf{x}) = |\frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{M}\sum_{j=N+1}^{N+M} x_j|$. We let $\pi(\mathbf{x})$ produces a new N+M sample that is a random permutation on the order of the elements in $\mathbf{x}_s$. There are (N+M)! possible permutations each of which is equally likely under H_0.

A simple introduction to permutation tests can be found in [1]. In [2] Joeseph P. Romano shows that for any finite set of transformations $\pi \in \Pi$ that are a mapping of $\mathbf{X}$ onto itself and for which $\pi(\mathbf{X})$ and $\mathbf{X}$ have the same distribution under H_0 the testing procedure described above produces a test of size α.

References

[1] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, 1994.

[2] J. P. Romano. On the behavoir of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, 1990.

[3] L. Scharf. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Publishing Company, 1991.

[4] L. Wasserman. *All of Statistics : A Concise Course in Statistical Inference*. Springer-Verlag New York, Inc., 2004.