1 Pointwise vs. Uniform Consistency in Level

Let $X_i, i = 1, \ldots, n$ be i.i.d. data from some distribution $P \in \mathbf{P}$. Suppose one wishes to test the null hypothesis $H_0 : P \in \mathbf{P}_0 \subsetneq \mathbf{P}$. To this end, one may consider a test function $\phi_n = \phi_n(X_1, \ldots, X_n)$ such that controls the probability of a Type 1 error in some sense. Ideally, we would like the test to satisfy

$$E_P[\phi_n] \leq \alpha \text{ for all } P \in \mathbf{P}_0 \ , \tag{1}$$

but many times this is too demanding a requirement. As a result, we may settle instead for tests such that

$$\limsup_{n \to \infty} E_P[\phi_n] \leq \alpha \text{ for all } P \in \mathbf{P}_0 \ . \tag{2}$$

Tests satisfying (1) are said to be of level α for $P \in \mathbf{P}_0$, whereas tests satisfying (2) are said to be pointwise asymptotically of level α for $P \in \mathbf{P}_0$. The hope is that if (2) holds, then (1) holds approximately, at least for large enough n. But this is not true. All that (2) ensures is that for each $P \in \mathbf{P}_0$ and $\epsilon > 0$ there is an $N(P)$ such that for all $n > N(P)$

$$E_P[\phi_n] \leq \alpha + \epsilon \ .$$

Importantly, the sample size required for the approximation to work, $N(P)$, may depend on P. As a result, it could be the case that for every sample size n (even, e.g., for $n = 10^{10}$) there could be $P = P_n \in \mathbf{P}_0$ such that

$$E_P[\phi_n] \gg \alpha \ .$$

Consider the following concrete example of this phenomenon. Suppose $\mathbf{P} = \{P \text{ on } \mathbf{R} : 0 < \sigma^2(P) < \infty\}$ and $\mathbf{P}_0 = \{P \in \mathbf{P} : \mu(P) = 0\}$. Let ϕ_n be the t-test; that is, $\phi_n = I\{\sqrt{n}\bar{X}_n > \hat{\sigma}_n z_{1-\alpha}\}$, where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution. We know that

$$E_P[\phi_n] \to \alpha \text{ for all } P \in \mathbf{P}_0 \ ,$$

but it turns out that the t-test suffers from the problem described above. In fact, we can show that for every $0 < c < 1$ and every sample size n there exists a $P = P_{n,c}$ such that

$$E_P[\phi_n] \geq c \ .$$

To see this, let n and c be given. Let P be the distribution that puts mass $1-p$ on p and mass p on $-(1-p)$. We will specify p in a minute, but first note that for such a distribution P all of the X_i are in fact equal to $p > 0$ with probability $(1-p)^n$. For such a sequence of observations, $\hat{\sigma}_n = 0$ and $\sqrt{n}\bar{X}_n > 0$, so $\phi_n = 1$. The probability of rejection, $E_P[\phi_n]$, is therefore at least $(1-p)^n$. Now all that remains is to choose p so that $(1-p)^n = c$; that is, $p = 1 - c^{1/n}$.

To rule this very disturbing possibility out, we need to ensure that the convergence in (2) is uniform for $P \in \mathbf{P}_0$; that is,

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_P[\phi_n] \leq \alpha \ . \tag{3}$$

Tests satisfying (3) are said to be uniformly asymptotically of level α for $P \in \mathbf{P}_0$. This requirement implies that for each $\epsilon > 0$ there is an N (which does not depend on P) such that for all $n > N$

$$E_P[\phi_n] \leq \alpha + \epsilon \ .$$

In the case of the t-test, the above example shows us that this is not true for $\mathbf{P} = \{P \text{ on } \mathbf{R} : 0 < \sigma^2(P) < \infty\}$ and $\mathbf{P}_0 = \{P \in \mathbf{P} : \mu(P) = 0\}$. But this does not mean that all hope is lost. Fortunately, the t-test does satisfy (3) for certain large classes of distributions that are somewhat smaller than $\mathbf{P}_0$.

2 Randomization Tests

The above discussion highlights the importance of requiring tests that behave well not just for each fixed P in some large class of distributions, but rather uniformly well over a large class of distributions. Of course, it goes without saying that whenever possible, we should seek tests that satisfy (1) for a large class of distributions. Typically, this is not possible, but for certain hypotheses it is. We will now discuss a general construction of such tests.

To this end, let X be distributed according to $P \in \mathbf{P}$ on a sample space $\mathcal{X}$. Suppose one wishes to test the null hypothesis $H_0 : P \in \mathbf{P}_0 \subsetneq \mathbf{P}$. Let $\mathbf{G}$ be a finite group of transformations of $\mathcal{X}$ onto itself $\mathcal{X}$. The following assumption, which we will refer to as the randomization hypothesis, allows for the construction of tests with the desired properties:

Assumption 2.1 *(Randomization Hypothesis)* For all $g \in \mathbf{G}$, X and gX have the same distribution whenever X has distribution $P \in \mathbf{P}_0$. ■

To help fix ideas, here are a few concrete examples of hypotheses that fit into this framework:

Example 2.1 Suppose $X_i, i = 1, \ldots, n$ be i.i.d real-valued random variables with distribution F, where F may be arbitrary. Here, $X = (X_1, \ldots, X_n)$. Suppose that under the null hypothesis F is symmetric about zero. For $i = 1, \ldots, n$, let ϵ_i take on either 1 or -1. Define a transformation g of $\mathbf{R}^n$ that by the rule that $x = (x_1, \ldots, x_n)$ is mapped to $(\epsilon_1 x_1, \ldots, \epsilon_n x_n)$ under g. Let $\mathbf{G}$ be the collection of the $M = 2^n$ such transformations. If X_i is distributed symmetrically about zero, then $\epsilon_i X_i$ and X_i have the same distribution. Since the X_i are independent, it follows that under the null hypothesis, gX and X have the same distribution. ■

Example 2.2 Suppose $Y_i, i = 1, \ldots, m$ are i.i.d observations with distribution P_Y and, independently, $Z_i, i = 1, \ldots, n$ are i.i.d. observations with distribution P_Z, where the distributions of P_Y and P_Z may be arbitrary. Here, $X = (Y_1, \ldots, Y_m, Z_1, \ldots, Z_n)$. Suppose that under the null hypothesis, $P_Y = P_Z$. Let π be a permutation of $1, \ldots, m + n$. Define a transformation g of $\mathbf{R}^{m+n}$ by the rule that $x = (x_1, \ldots, x_{m+n})$ is mapped to $(x_{\pi(1)}, \ldots, x_{\pi(m+n)})$ under g. Let $\mathbf{G}$ be the collection of the $M = (m + n)!$ such transformations. Under the null hypothesis, it is easy to see that gX and X have the same distribution. ■

Example 2.3 Suppose $(Y_i, Z_i), i = 1, \ldots, n$ are i.i.d. observations with distribution $P_{Y,Z}$, where $P_{Y,Z}$ is arbitrary. Let P_Y and P_Z denote the marginal distributions of $P_{Y,Z}$. Here, $X = ((Y_1, Z_1), \ldots, (Y_n, Z_n))$. Suppose that under the null hypothesis X_i and Y_i are independent. Define a transformation g of the sample space by the rule that $x = ((y_1, z_1), \ldots, (y_n, z_n))$ is mapped to $((y_1, z_{\pi(1)}), \ldots, (y_n, z_{\pi(n)}))$ under g. Under the null hypothesis, it is easy to see that gX and X have the same distribution. ■

This last example is of particular interest because of the following special case. Suppose it is desired to test whether some treatment has an impact on some outcome. Units are assigned at random to a treatment or a control group. Let D_i be an indicator variable for whether the ith unit was treated. Let W_i be the observed outcome for the ith unit. For example, D_i might be some medical treatment and W_i an indicator variable for mortality, or D_i might be a job training program and W_i an indicator variable for employment. We observe an i.i.d. sample of $(W_i, D_i), i = 1, \ldots, n$ and the null hypothesis specifies that W_i is independent of D_i. We may interpret this null hypothesis as one of no causal effect of D_i on W_i because the assignment to treatment is at random. It is important to understand why this rests upon the assumption that assignment to treatment is at random.

Remarkably, for each of these examples (and, more generally, for any testing problem in which the randomization hypothesis holds), we will be able to construct a test $\phi = \phi(X)$ of the null hypothesis such that

$$E_P[\phi] = \alpha \text{ for all } P \in \mathbf{P}_0 \ .$$

In order to describe the construction, let $T(X)$ be *any* real-valued test statistic for testing the null hypothesis. In Example 2.1, we may use $T(X) = |\bar{X}_n|$, whereas, in Example 2.2, we may use $T(X) = |\bar{Y}_m - \bar{Z}_n|$. Suppose we observe that $X = x$. Let $M = |\mathbf{G}|$ and denote by

$$T_{(1)}(x) \leq \cdots \leq T_{(M)}(x)$$

the ordered values of $T(gx)$ as g varies over $\mathbf{G}$. Let

$$\begin{aligned} k &= \lceil M(1-\alpha) \rceil = M - \lfloor M\alpha \rfloor \\ M^0(x) &= |\{1 \le j \le M : T_{(j)}(x) = T_{(k)}(x)\}| \\ M^+(x) &= |\{1 \le j \le M : T_{(j)}(x) > T_{(k)}(x)\}| \ . \end{aligned}$$

Let

$$a(x) = \frac{Mx - M^+(x)}{M^0(x)} \ .$$

Define

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > T_{(k)}(x) \\ a(x) & \text{if } T(x) = T_{(k)}(x) \\ 0 & \text{if } T(x) < T_{(k)}(x) \end{cases} \ .$$

Theorem 2.1 Suppose X has distribution $P \in \mathbf{P}$ on $\mathcal{X}$ and the problem is to test $H_0 : P \in \mathbf{P}_0$. Let $\mathbf{G}$ be a finite set of transformations of $\mathcal{X}$ onto $\mathcal{X}$. Suppose the Randomization Hypothesis holds. Given a test statistic $T(X)$, let ϕ be the test described above. Then,

$$E_P[\phi] = \alpha \text{ for all } P \in \mathbf{P}_0 \ .$$

Proof: By construction for every x,

$$\sum_{g \in \mathbf{G}} \phi(gx) = M^+(x) + a(x)M^0(x) = M\alpha \ .$$

Therefore,

$$M\alpha = E_P[\sum_{g \in \mathbf{G}} \phi(gX)] = \sum_{g \in \mathbf{G}} E_P[\phi(gX)] \ .$$

But, by the randomization hypothesis, $E_P[\phi(gX)] = E_P[\phi(x)]$. Hence,

$$M\alpha = \sum_{g \in \mathbf{G}} E_P[\phi(X)] = ME_P[\phi(X)] \ ,$$

from which the desired conclusion follows. ■

To gain further insight into why this works, let $\mathbf{G}_x = \{gx : g \in \mathbf{G}\}$. Because of the group structure of $\mathbf{G}$, these sets form a partition of the sample space $\mathcal{X}$. In other words, $\mathbf{G}_x \cap \mathbf{G}_{x'} = \emptyset$ for any $x \neq x'$ and $\bigcup_{x \in \mathcal{X}} \mathbf{G}_x = \mathcal{X}$. The Randomization Hypothesis says that under the null hypothesis the distribution of X conditional on $X \in \mathbf{G}_x$ is uniform on the set $\mathbf{G}_x$. Since this distribution does not depend on P, we can construct a test that is of level α conditional on $X \in \mathbf{G}_x$. The test therefore has the right size unconditionally as well.