1 Randomization Tests

Our earlier 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 are of level α 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 1.1 *(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 1.1 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 1.2 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 1.3 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

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

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 1.1 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.

One can also construct p-values for randomization tests as follows:

$$\hat{p} = \frac{1}{M} \sum_{g \in \mathbf{G}} I\{T(gX) \geq T(X)\} \ .$$

It can be shown that under the null hypothesis

$$P\{\hat{p} \leq u\} \leq u \text{ for all } 0 \leq u \leq 1 \ .$$

Because $\mathbf{G}$ may be large, one may need to resort to a stochastic approximation to the randomization test described above. For example, let $g_i, i = 1, \ldots, B-1$ be i.i.d. with the uniform distribution over $\mathbf{G}$ and let g_B be equal to the identity transformation. It can be shown that

$$\tilde{p} = \frac{1}{B} \sum_{1 \leq i \leq B} I\{T(g_i X) \geq T(X)\}$$

also satisfies

$$P\{\tilde{p} \leq u\} \leq u \text{ for all } 0 \leq u \leq 1 \ .$$

2 Asymptotic Behavior of Randomization Tests

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$, where $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with finite, nonzero variance. Suppose we are interested in testing whether the $P \in \mathbf{P}_0 = \{P \in \mathbf{P} : \mu(P) = 0\}$ versus $P \in \mathbf{P}_1 = \{P \in \mathbf{P} : \mu(P) > 0\}$. If we had assumed that the underlying distribution were symmetric, we could consider using a randomization test, as described above. Remarkably, even if the underlying distribution is not symmetric, the randomization test still results in a test that is pointwise asymptotically of level α for $P \in \mathbf{P}_0$. The advantage of such a test is obvious: it will be of level α for those $P \in \mathbf{P}_0$ that are symmetric. On the other hand, such an advantage may come at a cost in terms of power. For example, if P were normally distributed, then one may want to consider using the optimal t-test. Remarkably, it is possible to show that in a certain sense the loss of power of the randomization test relative to the t-test is small, at least for large sample sizes and a suitable choice of test statistic.

In order to answer the question described above, we must first develop some tools to analyze the large-sample behavior of randomization tests. It is useful now to index our earlier notation by the sample size n. To that end, let X^n be distributed according to P_n on a sample space $\mathcal{X}_n$. Typically, $X^n = (X_1, \ldots, X_n)$. If the data consists of n i.i.d. observations, then $P_n = P^n$, the n-fold product of the distribution of a single observation. Let $\mathbf{G}_n$ be a finite group of transformations from $\mathcal{X}_n$ onto itself and denote by $T_n(X^n)$ the test statistic of interest. Let $\hat{R}_n(t)$ be the randomization distribution of T_n defined by

$$\hat{R}_n(t) = \frac{1}{M_n} \sum_{g \in \mathbf{G}_n} I\{T_n(gX^n) \leq t\} ,$$

where $M_n = |\mathbf{G}_n|$. Let $\hat{r}_n(1-\alpha)$ be the $1-\alpha$ quantile of $\hat{R}_n(t)$, i.e.,

$$\hat{r}_n(1-\alpha) = \inf\{t \in \mathbf{R} : \hat{R}_n(t) \geq 1-\alpha\} .$$

(Note that in the notation of the previous section, $\hat{r}_n(1-\alpha)$ is simply

$T_{(k)}(X)$.) In this notation, the randomization test of the previous section can be described as the test that rejects when

$$T_n(X^n) > \hat{r}_n(1-\alpha) \ ,$$

rejects with some probability when $T_n(X^n) = \hat{r}_n(1-\alpha)$, and doesn't reject otherwise. For simplicity, assume that the test doesn't reject in the event that $T_n(X^n) = \hat{r}_n(1-\alpha)$.

The key to analyzing the asymptotic behavior of the randomization test therefore lies in the asymptotic behavior of $\hat{r}_n(1-\alpha)$, which is in turn determined by the asymptotic behavior of $\hat{R}_n(t)$. The following theorem, due to Hoeffding (1952), is useful for this purpose. Note that the result does not assume that the Randomization Hypothesis holds.

Theorem 2.1 Suppose X^n has distribution P_n on $\mathcal{X}_n$. Let G_n be a finite group of transformations from $\mathcal{X}_n$ onto itself. Let G_n and G_n' be uniformly distributed over $\mathbf{G}_n$. Suppose G_n, G_n' and $\mathcal{X}_n$ are mutually independent. If

$$(T_n(G_n X^n), T_n(G_n' X^n)) \xrightarrow{d} (T, T')$$

under P_n, where T and T' are independent with common c.d.f. $R(t)$, then

$$\hat{R}_n(t) \xrightarrow{P} R(t) \ ,$$

under P_n for every continuity point t of R. If R is continuous and strictly increasing at $r(1-\alpha) = \inf\{t \in \mathbf{R} : \hat{R}(t) \geq 1-\alpha\}$, then

$$\hat{r}_n(1-\alpha) \xrightarrow{P} r(1-\alpha) \ .$$

Proof: Let t be a continuity point of R. Then, by the assumed convergence in distribution,

$$E_{P_n}[\hat{R}_n(t)] = P_n\{T_n(G_n X^n) \leq t\} \to R(t) \ .$$

It therefore suffices to show that

$$E_{P_n}[\hat{R}_n^2(t)] \to R^2(t) \ ,$$

which implies that $\mathrm{Var}_{P_n}[\hat{R}_n(t)] \to 0$. Note that

$$\begin{aligned} E_{P_n}[\hat{R}_n^2(t)] &= \sum_{M_n^2} \sum_{g \in \mathbf{G}_n} \sum_{g' \in \mathbf{G}_n} P_n\{T_n(gX^n) \le t, T_n(g'X^n) \le t\} \\ &= P_n\{T_n(G_nX^n) \le t, T_n(G_n'X^n) \le t\} \\ &\to R^2(t) \ , \end{aligned}$$

by the assumed convergence in distribution. Hence, $\hat{R}_n(t) \xrightarrow{P} R(t)$ under P_n. The convergence of $\hat{r}_n(1-\alpha)$ now follows from earlier arguments. ■

We now apply this result to the situation described at the beginning of the section. Consider using the randomization test based on symmetry of the underlying distribution (which we are not assuming):

$$\phi_{1,n} = I\{T_n > \hat{r}_n(1-\alpha)\} \ ,$$

where

$$T_n = \sqrt{n}\bar{X}_n \ .$$

We first establish that $\phi_{1,n}$ is pointwise asymptotically of level α for $P \in \mathbf{P}_0$. To this end, consider any $P \in \mathbf{P}_0$ and $\epsilon_1, \ldots, \epsilon_n, \epsilon_1', \ldots, \epsilon_n'$ be an i.i.d. sequence of random variables that put equal weight on 1 and -1. We verify the conditions of the preceding theorem with $R(t) = \Phi(t/\sigma(P))$. We must determine the limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (\epsilon_i X_i, \epsilon_i' X_i) \ .$$

Note that

$$E_P[\epsilon_i X_i] = E_P[\epsilon_i' X_i] = 0$$

and

$$E_P[(\epsilon_i X_i)^2] = E_P[(\epsilon_i' X_i)^2] = \sigma^2(P) \ .$$

Moreover,

$$\mathrm{Cov}_P[\epsilon_i X_i, \epsilon_i' X_i] = 0 \ .$$

Hence, by the usual central limit theorem,

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (\epsilon_i X_i, \epsilon_i' X_i) \xrightarrow{d} (T, T') \ ,$$

where T and T' are independent and each distributed as $N(0, \sigma^2(P))$. Hence, by Theorem 4.1, $\hat{R}_n(t) \xrightarrow{P} R(t)$ for each $t \in \mathbf{R}$ and $\hat{r}_n(1-\alpha) \xrightarrow{P} \sigma(P) z_{1-\alpha}$. Therefore,

$$E_P[\phi_{1,n}] \to \alpha$$

for any such P.

We now examine the power of $\phi_{1,n}$ against a sequence of alternatives of the form $P_n = N(h/\sqrt{n}, \sigma^2)$. By the above, under $P_0 = N(0, \sigma^2)$, $\hat{r}_n(1-\alpha) \xrightarrow{P} \sigma(P) z_{1-\alpha}$. By contiguity, we have immediately that under P_n $\hat{r}_n(1-\alpha) \xrightarrow{P} \sigma(P) z_{1-\alpha}$ as well. Since $T_n \sim N(h, \sigma^2)$ under P_n, we have that

$$E_{P_n}[\phi_{1,n}] \to 1 - \Phi(z_{1-\alpha} - \frac{h}{\sigma}) \ .$$

Remarkably, this is also the limiting power against such alternatives for the t-test as well. In this sense, the loss of power of the randomization test relative to the t-test is small, at least for large sample sizes.