Even More Notes on Testing

Dealing with Composite Hypotheses

- The Neyman-Pearson Lemma solves the problem of how to construct best tests of a single null hypothesis against a single alternative hypothesis. In applications, however, the hypotheses of interest rarely exactly specify the distribution of the observations.

- We saw an example (a test of $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ with observation $X \sim \mathcal{N}(\mu, 1)$) where a single test is best against all (one-sided) alternatives, making the test uniformly most powerful. In that instance, it doesn’t matter that the alternative is composite.

- But in many other examples of interest, also the null hypothesis is composite, and/or no such uniformly most powerful test exists. The following notes discuss standard strategies how to deal with the composite nature of hypotheses. In complicated testing problems, several of these techniques can be combined, although we will consider them one-by-one.

Invariance

- Suppose we observe $X \sim \mathcal{N}(\mu, 1)$, and we want to test the null hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. We know that no uniformly most powerful test exists for this problem: The best test against the alternative $\mu = -1$, say, rejects for small values of $X$, and the best test against the alternative $\mu = 1$ rejects for large values of $X$.

- The hypothesis problem we are facing exhibits a certain symmetry: If instead of observing $X$, we were to observe $X^* = -X$, we would face the exact same problem: With $\mu^* = E[X^*] = -\mu$, we have $H_0^* : \mu^* = 0$ and $H_1^* : \mu^* \neq 0$. There is no difference whatsoever between the starred problem and the original problem.

- This suggests that if we observe $X = 3$ and decide to reject, then we should also reject for the observation $X = -3$. Formally, one might demand that given the symmetry in the problem, a test satisfies

$$\varphi(x) = \varphi(-x).$$

A test with that property is called invariant to the transformation $x \rightarrow -x$.

- More generally, for a general $n \times 1$ observation $X$, consider a group of transformations $G = \{g_a | a \in A\}$, where $g_a : \mathbb{R}^n \mapsto \mathbb{R}^n$. In the example, we could set $A = \{-1, 1\}$, with
The question thus becomes how to determine the best invariant test, that is the most powerful test of level $\alpha$ in the class of all tests that satisfy (1). There is a powerful theorem that describes all invariant tests in a very tractable way, which relies on the concept of a maximal invariant:

Definition: A function $h : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is said to be a maximal invariant with respect to the group of transformations $\mathcal{G}$ if

(i) $h(g(x)) = h(x)$ for all $g \in \mathcal{G}$ (invariance)

(ii) $h(x) = h(x^*)$ implies $x = g(x^*)$ for some $g \in \mathcal{G}$ (maximality).

Example: Let $h(x) = |x|$. Then $h(x)$ is a maximal invariant, since (i) $|−x| = |x|$ and (ii) when $|x| = |x^*|$, then either $x = g_1(x^*) = x^*$, or $x = g_{−1}(x^*) = −x^*$. Note that maximal invariants are not unique: $h_2(x) = x^2$ is also a maximal invariant. (Check!)

Theorem: Let $h(x)$ be a maximal invariant with respect to $\mathcal{G}$. Then a necessary and sufficient condition for the test $\varphi(x)$ to be invariant is that it depends on $x$ only through $h(x)$, that is there exists a function $\vartheta$ for which $\varphi(x) = \vartheta(h(x))$ for all $x$.

Sketch of proof: Sufficiency is clear, since a maximal invariant is invariant: $\varphi(g(x)) = \vartheta(h(g(x))) = \vartheta(h(x)) = \varphi(x)$. For necessity, note that what could go wrong is that there exist $x$ and $x^*$ for which $h(x) = h(x^*)$, yet $\varphi(x) \neq \varphi(x^*)$. But by maximality, if $h(x) = h(x^*)$ then $x = g(x^*)$ for some $g \in \mathcal{G}$, so that also $\varphi(x) = \varphi(x^*)$.

The Theorem tells us that if we restrict attention to invariant tests, we can treat $h(X)$ as the effective observation, since all invariant tests can be written as a function of $h(X)$. In the example, the pdf of $Y = h(X) = |X|$ is

\[ f_Y(y, \mu) = f_X(y, \mu) + f_X(-y, \mu) \quad \text{for } y \geq 0 \text{ and } 0 \text{ otherwise} \]

where $f_X(x, \mu)$ is the pdf of a normal with mean $\mu$ and variance 1. Note that $f_Y(y, \mu) = f_Y(y, |\mu|)$. Consider the best test of $H_0 : \mu = 0$ against $H_1 : \mu = \mu_1 \neq 0$ based on $h(X) = |X|$ (which by the Theorem, is the best invariant test). By the Neyman-Pearson Lemma, this test rejects for large values of $f_Y(Y, \mu_1)/f_Y(Y, 0)$, and a calculation shows that this is equivalent to rejecting for large values of $Y = h(X) = |X|$.
• Invariance can also be useful to eliminate nuisance parameters that appear under both hypotheses: Let $X_i \sim \text{iid } \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, and we want to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. One might consider tests that are invariant to the group of transformations

$$\{X_i\}_{i=1}^{n} \rightarrow \{cX_i\}_{i=1}^{n} \text{ for } c > 0.$$ 

A maximal invariant is given by

$$h(X) = \begin{cases} 
0 \text{ if } X_1 = X_2 = \cdots = X_n \\
\{X_i/s\}_{i=1}^{n} \text{, where } s^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \text{ otherwise} 
\end{cases}$$

and it is very painful (but not conceptually hard) to show that, by the Neyman-Pearson Lemma, the uniformly most powerful invariant test rejects for large values of $t = \bar{X}/s$.

• A generalization of the first example: Suppose we observe the $k \times 1$ vector $X \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma$ known, and we want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. This is a composite alternative. Now consider the group of transformations

$$x \rightarrow \Sigma^{1/2} O \Sigma^{-1/2} x \text{ for } O \text{ an orthogonal matrix.} \quad (2)$$

Recall that a matrix is orthogonal iff $O'O = I$. A maximal invariant to this group is $h(x) = x'\Sigma^{-1} x$. Invariance is clear. For maximality, note that for any $v \in \mathbb{R}^k$, we can always find $O$ such that $Ov$ is zero in the last $k-1$ columns (simply make the first row of $O$ proportional to $v$), and $||Ov|| = ||v||$. Thus, if $x' \Sigma^{-1} x = x'^* \Sigma^{-1} x^*$, we can find $O$ and $O^*$ such that $O \Sigma^{-1/2} x = O^* \Sigma^{-1/2} x^*$ and the last $k-1$ rows are equal to zero. Hence, $x = \Sigma^{1/2} O'O^* \Sigma^{-1/2} x^* = \Sigma^{1/2} O_1 \Sigma^{-1/2} x^*$ for the orthonormal matrix $O_1 = O'O^*$.

The distribution of $h(X) = X' \Sigma^{-1} X$ is $\chi^2_k$ under the null hypothesis, and non-central chi-squared under $H_1$. The Neyman-Pearson Lemma is applicable, and it turns out that we should reject for large value of $h(X) = X' \Sigma^{-1} X$ for any value of $\mu$. Thus, rejecting for large values of $h(X) = X' \Sigma^{-1} X$ is the uniformly most powerful invariant (to (2)) test.

• In applications of invariance, the real question is whether it makes sense to impose a particular invariance requirement. This depends on the application.
Unbiasedness

• Consider again the problem of testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$, based on a single observation $X \sim \mathcal{N}(\mu, 1)$. As noted before, the best test against the alternative $\mu = 1$ rejects for large values of $X$. To be specific, the best 5% level test rejects if $X > 1.645$, because $P(X > 1.645|H_0) = P(Z > 1.645) = 0.05$. Note that this test is biased: for $\mu = -1$, its rejection probability is $P(X > 1.645|\mu = -1) = P(Z - 1 > 1.645) < 0.05$. This might be considered undesirable, because there are values under the alternative that systematically lead to a rejection probability below the level of test.

• Definition: A test of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is unbiased level $\alpha$ if

$$\pi(\theta) = E_{\theta}[\varphi(X)] \geq \alpha \text{ for all } \theta \in \Theta_1.$$

• When testing a single null hypothesis $\theta = \theta_0$, this requires that the power function $\pi(\theta)$ is minimized at $\theta_0 = 0$. If $\pi$ is differentiable, this implies

$$\frac{d\pi}{d\theta}(\theta_0) = 0.$$

• Normal Example: For any test $\varphi$

$$\pi(\mu) = E_{\mu}[\varphi(X)] = \int \varphi(x)f(x, \mu)dx$$

where $f(x, \mu) = (2\pi)^{-1/2} \exp[-\frac{1}{2}(x - \mu)^2]$. Differentiating under the integral yields

$$\frac{d\pi}{d\mu} = \int \varphi(x)f(x, \mu)(x - \mu)dx$$

so that $\frac{d\pi}{d\mu}(0) = 0$ yields $E_0[\varphi(X)X] = 0$. The best unbiased test against $\mu = \mu_1 \neq 0$ thus maximizes

$$E_{\mu_1}[\varphi(X)] = \int \varphi(x)f(x, \mu_1)dx$$

s.t. $E_0[\varphi(X)] = \int \varphi(x)f(x, 0)dx = \alpha$ and $E_0[\varphi(X)X] = \int \varphi(x)f(x, 0)x dx = 0$.

Set up Lagrangian

$$\int \varphi(x)f(x, \mu_1)dx - \lambda_0 \int \varphi(x)f(x, 0)dx - \lambda_1 \int \varphi(x)f(x, 0)x dx$$

$$= \int \varphi(x)f(x, \mu_1)dx - \lambda_0 \int \varphi(x)[f(x, 0) + \frac{\lambda_1}{\lambda_0}f(x, 0)x]dx$$

4
This is the same problem that we faced in the Neyman-Pearson derivation, with a 'null density' equal to $f(x, 0) + \frac{\lambda_1}{\lambda_0} f(x, 0) x$. By the same logic, the form of the optimal test is thus

$$
\varphi(x) = \begin{cases} 
1 & \text{if } f(x, \mu_1) > \lambda_0 (f(x, 0) + \frac{\lambda_1}{\lambda_0} f(x, 0) x) \\
0 & \text{otherwise}
\end{cases}
$$

A calculation now shows (check!) that the best unbiased tests rejects for large values of $|X|$.

- Unbiasedness is a constraint that is mostly employed to argue for the reasonableness of particular two-sided tests (that is, in tests with an alternative that doesn’t specify whether the parameter of interest is smaller or larger than the hypothesized value). In complicated testing problems involving nuisance parameters, no (nontrivial) unbiased test may exist.

**Weighted Average Power**

- Let $X$ have density $f(x, \theta)$, with both $x$ and $\theta$ possibly vectors. Consider a test of a simple null hypothesis against a composite alternative

$$
H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1.
$$

The power of a test $\varphi$ is then given by

$$
\pi(\theta) = E_\theta[\varphi(X)] = \int \varphi(x) f(x, \theta) dx.
$$

In absence of a uniformly most powerful test, no single test $\varphi$ maximizes $\pi(\theta)$ for all values of $\theta$.

- Example: Let $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu$ is $k \times 1$, and the testing problem is $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. No uniformly most powerful test exists here: Let $\mu_1 \neq 0$ with $\mu_1 \in \mathbb{R}^k$. The application of the NP Lemma shows that the best test of $H_0 : \mu = 0$ against $H_1 : \mu = \mu_1$ rejects for large values of

$$
LR_c(\mu_1) = \exp[\mu_1 \Sigma^{-1} X - \frac{1}{2} \mu_1' \Sigma^{-1} \mu_1]
$$

or, equivalently, $\mu_1' \Sigma^{-1} X$, so the critical region depends on $\mu_1$. Note that an equivalent test statistic is

$$
\frac{\mu_1' \Sigma^{-1} X}{\sqrt{\mu_1' \Sigma^{-1} \mu_1}} \sim \mathcal{N}(\sqrt{\mu_1' \Sigma^{-1} \mu_1}, 1).
$$

The power of this 'point-optimal' test depends only on $\mu_1' \Sigma^{-1} \mu_1$. In other words, alternatives $\{\mu_1 : \mu_1' \Sigma^{-1} \mu_1 = \lambda\}$ are equally difficult to distinguish from the null hypothesis.
Since no uniformly most powerful test exists, one might instead seek to maximize some average of the power function. Let $w(\theta)$ be the probability density function of a random variable that takes on values in $\Theta_1$. Then weighted average power is equal to

$$WAP = \int \pi(\theta)w(\theta)d\theta = \int \left( \int \varphi(x)f(x,\theta)dx \right) w(\theta)d\theta = \int \varphi(x)\left( \int w(\theta)f(x,\theta)d\theta \right) dx$$

where the interchange of the order of integration is allowed, because all integrands are nonnegative. Note that $w_{mix}(\theta) = \int f(x,\theta)d\theta$ is again a probability density: it is nonnegative, and it integrates to one, since $
abla \int w_{mix}(x)dx = \int w(\theta)f(x,\theta)d\theta dx = \int w(\theta)\int f(x,\theta)dx d\theta = 1$. Thus, a test that maximizes WAP equivalently maximizes power against the single alternative $H_{1}^{mix}$: $X$ has density $f_{mix}(x) = \int w(\theta)f(x,\theta)d\theta$,

And by the NP Lemma, the best test rejects for large values of $f_{mix}(x) = \frac{f(x,\theta)}{f(x,\theta_0)} = \int w(\theta)LR_c(\theta)d\theta$

where $LR_c(\theta) = f(x,\theta)/f(x,\theta_0)$.

Example ctd: Let $w$ be the pdf of $N(0,c\Sigma)$. The best rejects for large values of

$$\int LR_c(\mu)w(\mu)d\mu = |c\Sigma|^{-1/2} \int \exp[X'S^{-1}\mu - \frac{1}{2}\mu'\Sigma^{-1}\mu - \frac{1}{2}c^{-1}\mu'\Sigma^{-1}\mu]d\mu = |c\Sigma|^{-1/2} \int \exp[-\frac{1}{2}(\mu - \Sigma^{-1}\Omega^{-1}X)'\Omega(\mu - \Sigma^{-1}\Omega^{-1}X) + \frac{1}{2}X'S^{-1}\Omega^{-1}\Sigma^{-1}X]d\mu = |c\Sigma|^{-1/2}|\Omega|^{-1/2} \exp[\frac{1}{2}X'S^{-1}\Omega^{-1}\Sigma^{-1}X]$$

where $\Omega = (1 + c^{-1})\Sigma^{-1}$, so that the weighted average power maximizing test rejects for large values of $X'S^{-1}X$ (no matter what the value of $c$ is, so there exists a UMP test with respect to $c$). Note that the weighting function $w(\mu)$ puts equal weight
on the elements in the set \( \{ \mu_1 : \mu_1 \Sigma^{-1} \mu_1 = \lambda \} \) for all \( \lambda \). It can be shown that for any weight function that satisfies this (i.e. puts equal weight on alternatives that are equally difficult to distinguish from the null hypothesis), the optimal test rejects for large values of \( X' \Sigma^{-1} X \).

- A focus on weighted average power makes the alternative hypothesis effectively simple. The question is what weighting function one should use. Note that in the above example, the choice of putting equal weight on alternatives that are equally difficult to distinguish from the null hypothesis is purely 'statistically' motivated—it might be that in a particular problem, we care much more about deviations in one direction than about others.

**Least Favorable Distributions**

- Let \( X \) be a random vector. Consider a test of a composite null hypothesis against a simple alternative

  \[
  H_0 : \text{the density of } X \text{ is } f(x, \theta), \theta \in \Theta_0 \\
  H_1 : \text{the density of } X \text{ is } g(x).
  \]

- Theorem: Suppose \( \Lambda \) is a probability distribution with support equal to a subset of \( \Theta_0 \).

  (a) Let \( \varphi \) be any level \( \alpha \) test under \( H_0 \), i.e. \( \sup_{\theta} \int \varphi(x) f(x, \theta) dx \leq \alpha \). For any \( \Lambda \), the best level \( \alpha \) test \( \varphi_\Lambda \) of

  \[
  H_\Lambda : \text{the density of } X \text{ is } f_\Lambda(x) = \int f(x, \theta) d\Lambda(\theta)
  \]

  against \( H_1 \) has at least as much power than \( \varphi \).

  (b) If \( \Lambda = \Lambda^* \) is such that \( \varphi_{\Lambda^*} \) is also of level \( \alpha \) under \( H_0 \), then \( \varphi_{\Lambda^*} \) is the best level \( \alpha \) test of \( H_0 \), and \( \Lambda^* \) is called the least favorable distribution.

- Proof: (a) The test \( \varphi \) is also of level \( \alpha \) under \( H_\Lambda \), since \( \int \varphi(x) f_\Lambda(x) dx = \int \int \varphi(x) f(x, \theta) dxd\Lambda(\theta) \leq \sup_{\theta} \int \varphi(x) f(x, \theta) dx \leq \alpha \). But \( \varphi_\Lambda \) is the best level \( \alpha \) test of \( H_\Lambda \) against \( H_1 \), so its power is at least as large as that of \( \varphi \).

  (b) Immediate from part (a).

- Example: Let \( X \sim \mathcal{N}(\mu, 1) \), and \( H_0 : \mu \leq 0 \) against \( H_1 : \mu = \mu_1 > 0 \). Let \( \Lambda \) be the distribution that puts all its mass on \( \mu = 0 \), so that \( H_\Lambda \) becomes \( H_\Lambda : X \sim \mathcal{N}(0, 1) \).
The best level $\alpha$ test of $H_0$ against $H_1$ is given by $\varphi_\Lambda(x) = 1[x > c]$, where $c$ solves $P(Z > c) = \alpha$, $Z \sim \mathcal{N}(0, 1)$. Now $\varphi_\Lambda(x)$ is also of level $\alpha$ under $H_0$, since for all $\mu \leq 0$, \[ \int \varphi_\Lambda(x)f(x, \mu)dx = P(Z - \mu > c) = P(Z > c + \mu) \leq \alpha. \] Thus, by the Theorem, $\Lambda$ is the least favorable distribution, and $\varphi_\Lambda(x) = 1[x > c]$ is the best test.

- The name least favorable comes from (a), the fact that $\Lambda^*$ induces the lowest possible power. There is no systematic way of deriving $\Lambda^*$. Rather, one guesses a candidate $\Lambda^*$ and shows that it is in fact least favorable. This can be very hard.

- Sophisticated example: Suppose $X \sim \mathcal{N}(\mu, \Sigma)$, with $\Sigma$ known. For a given vector $v$, let $\beta = v'\mu$, and we are interested in testing $H_0 : \beta = 0$ against $H_1 : \beta > 0$. Claim: the best test rejects for large values of $v'X$.

Proof: Fix arbitrary $\mu_1$ with $v'\mu_1 > 0$, and consider the simple alternative $H^s_1(\mu_1) : \mu = \mu_1$. Define
\[
\mu_0 = \mu_1 - \frac{v'\mu_1}{v'\Sigma v} \Sigma v
\]
and note that $v'\mu_0 = 0$. Consider the simple null hypothesis $H^s_0(\mu_0) : \mu = \mu_0$ (the point mass at $\mu_0$ will turn out to be the least favorable distribution for the alternative $\mu_1$). By the NP Lemma, the best test of $H^s_0(\mu_0)$ against $H^s_1(\mu_1)$ rejects for large values of
\[
X'\Sigma^{-1} \mu_1 - X'\Sigma^{-1} \mu_0 = \frac{v'\mu_1}{v'\Sigma v} v'X
\]
or, equivalently, for large values of $v'X$. This test does not depend on $\mu_1$, and $\mu_1$ was arbitrary (in $H_1$), so $v'X$ is the uniformly most powerful test of $H_0 : v'\mu = 0$ against $H_1 : v'\mu > 0$. 