STA 732: INFERENCE

Notes 2. Neyman-Pearsonian Classical Hypothesis Testing

B&D 4

1 Testing as a “rule”

Fisher’s quantification of “extremeness of observed evidence” clearly lacked rigorous mathematical interpretation. Neyman and Pearson sought to find a logically coherent interpretation of Fisher’s significance testing and developed the modern classical paradigm based on “long run frequentist guarantees of statistical rules”.

Under the (Neyman-Pearsonian) classical setting, hypothesis testing is set up as a comparative assessment of two assertions: $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ and given observations $X = x$, one needs to take one of the two decisions: “accept $H_0$” or “accept $H_1$ (i.e., reject $H_0$)” according to some rule of decision making. If $T(X)$ is a statistic\(^1\) that measures evidence against $H_0$ in data $X$, then one can create a decision rule out of it by first deciding how large $T(X)$ needs to be to convince us of enough evidence against $H_0$. That is, we could fix a “critical value” $c$ and define the testing rule:

$$\text{reject } H_0 \text{ if } T(X) \geq c$$

Clearly any number rules could be created by considering different critical values. Also, we have complete freedom in choosing the test statistic. Even more generally, if $S$ denotes the sample space of all possible data $X$, then any test rule is given by

$$\text{reject } H_0 \text{ if } X \in R$$

for some subset $R \subset S$, and the converse is also true.

When a testing rule is applied to data $X$, one of the 4 situations must arise:

1. $H_0$ is true and we reject $H_0$: a type I error;
2. $H_0$ is true and we accept $H_0$: a correct decision;
3. $H_1$ is true and we reject $H_0$: a correct decision;
4. $H_1$ is true and we accept $H_0$: a type II error.

Therefore the long run behavior of the rule could be mathematically quantified by its propensity to commit type I and type II errors. The big idea behind the N-P approach was to use these quantifications as the mathematical cornerstone of statistical practice. Notice that the quantifications apply only to the rules, and are completely divorced from the application of any rule to the actual observed data. We will come back to this point below.

\(^1\)Formally, a “statistic” is any summary of the data $X$
2 Fixed level testing

It is clear that for any given test statistic $T(X)$, it is not possible to choose a critical value $c$ that will simultaneously minimize the two types of error probabilities. Take for example of testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ based on a single observation $X \sim \text{Normal}(\mu, 1)$ where $\mu_0 < \mu_1$ are given numbers. Then $P(\text{type I error}) = 1 - \Phi(c - \mu_0)$ and $P(\text{type II error}) = \Phi(c - \mu_1)$ where $\Phi(z)$ denotes the cdf of the standard normal distribution. Increasing $c$ will reduce the type I error probability but increase the type II error probability, and decreasing it will have the opposite effect (Figure 1). A more degenerate situation arises in a more practically relevant setting of testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ for some given number $\mu_0$. In this case $\max P(\text{type I error}) = 1 - \Phi(c - \mu_0) = 1 - \max P(\text{type II error})$. Therefore a clear choice of $c$ is not possibly unless a distinction is made between the two hypotheses.

The N-P testing paradigm places much more emphasis on $H_0$ than on $H_1$, with the philosophy that “no data” means “no evidence against $H_0$” and the primary long-run characteristic of a testing rule is its “size” defined as:

\[
\text{size} = \max_{\theta \in \Theta_0} P(T(X) \geq c|\theta) 
\]

or more generally, $\text{size} = \max_{\theta \in \Theta_0} P(X \in R|\theta)$ if the rule is written as “reject $H_0$ if $X \in R$”.

N-P’s “fixed level testing” has the following recipe:

1. Carefully choose the labels $H_0$ and $H_1$ for the two assertions about $\theta$; $H_0$ should denote the status-quo case, i.e., the assertion you will accept if no data was available.

2. Identify a “good” test statistic $T(X)$ quantifying evidence against $H_0$.

3. Fix a small nominal level of significance $\alpha$ (typically $\alpha = 5\%$, 1\% or 10\%).
4. Find the critical value \( c = c(\alpha) \) for which the test rule “reject \( H_0 \) if \( T(X) \geq c \)” has size \( \alpha \) or smaller (as close to \( \alpha \) as possible without exceeding it).

5. Use the rule “reject \( H_0 \) if \( T(X) \geq c(\alpha) \)” as a (fixed) \( \alpha \) level rule.

Some comments are due. Consider testing \( H_0 : \mu \leq 0 \) vs. \( H_1 : \mu > 0 \) based on a single observation \( X \sim \text{Normal}(\mu, 1) \). The size of any rule “reject \( H_0 \) if \( X \geq c \)” equals \( \max_{\mu \leq 0} P(X \geq c | \mu) = P(X \geq c | \mu = 0) = 1 - \Phi(c) \). So the size 5\% rule is given by “reject \( H_0 \) if \( X \geq 1.64 \)”.

Note that the rule rejects \( H_0 \) both for \( X = 2 \) and \( X = 10 \) (at level 5\%) and offers nothing to differentiate the two cases. Fisher’s significance testing approach, on the other hand will produce p-value \( = P(X \geq 2 | \mu = 0) = 0.022 \) when \( X = 2 \) and p-value \( = P(X \geq 10 | \mu = 0) = 7 \times 10^{-24} \) when \( X = 10 \) conveying a much stronger evidence against \( H_0 \) in the latter. The failure of N-P’s approach to offer any distinction between the two cases was heavily criticized by Fisher.

A second comment is that to carry out a level \( \alpha \) test using a test statistic \( T(X) \) is equivalent to the rule “reject \( H_0 \) if p-value based on \( T < \alpha \)”.

3 Choice of test statistic

Although N-P’s approach fails provide application specific interpretation, it offers a very powerful theoretical platform to answer many questions related to Fisher’s significance testing, e.g., what test statistic \( T(X) \) should one use? Fisher did not have a convincing answer to this (essentially said a serious scientist knows what to use), though his LR\((X)\) statistic will turn out to be a correct choice under the N-P paradigm.

In order to choose between two test statistics \( T_1(X) \) and \( T_2(X) \) to carry out level \( \alpha \) testing, N-P advocated looking at the size-\( \alpha \) rules based on these statistics: “reject \( H_0 \) if \( T_1(X) \geq c_1(\alpha) \)” and “reject \( H_0 \) if \( T_2(x) \geq c_2(\alpha) \)” and using the one with lower type II error probabilities. To facilitate the discussion, let’s introduce the notion of “power function” of a test rule “reject \( H_0 \) if \( T(X) \geq c \)”:

\[
pow(\theta) := P(T(X) \geq c | \theta), \ \theta \in \Theta
\]

or more generally, \( pow(\theta) := P(X \in R | \theta) \) if the rule is given as “reject \( H_0 \) if \( X \in R \)”.

Clearly, for any \( \theta \in \Theta_0 \), \( pow(\theta) = P(\text{type I error} | \theta) \) and the size of the test is precisely \( \max_{\theta \in \Theta_0} pow(\theta) \). Also, for any \( \theta \in \Theta_1 \), \( pow(\theta) = 1 - P(\text{type II error} | \theta) \). Therefore N-P’s suggestion is equivalent to “for any two tests with identical size, use the one with more power at the alternative”. The best test along this line, if it exists, is called the uniformly most powerful (UMP) test, defined as follows.

**Definition 1 (UMP test).** A test rule \( \delta \) is called the UMP test at level \( \alpha \) if size \( \delta \leq \alpha \) and any other testing rule \( \tilde{\delta} \) with size \( \tilde{\delta} \leq \alpha \) satisfies: \( pow(\delta; \theta) \geq pow(\tilde{\delta}; \theta) \) for all \( \theta \in \Theta_1 \).

N-P not only offered a theoretical definition of the “optimal test”, they also showed how to construct one using the likelihood ratio statistic in the case of testing a simple null against a simple alternative\(^2\).

\(^2\) \( H_0 : \theta \in \Theta_0 \) is called “simple” if \( \Theta_0 = \{\theta_0\} \) is a singleton set, otherwise \( H_0 \) is called a composite hypothesis. Same applies to \( H_1 \).
Lemma 1 (Neyman-Pearson). Consider testing \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta = \theta_1 \) based on an \( X \sim f(x|\theta) \). For any \( c > 0 \), the testing rule

\[
\text{"reject } H_0 \text{ if } \Lambda(X) = \frac{f(X|\theta_1)}{f(X|\theta_0)} \geq c
\]

is UMP at level \( \alpha = \alpha(c) = P(\Lambda \geq c|\theta_0) \).

**Proof.** Let \( \mathcal{R} = \{ x : f(x|\theta_1) \geq cf(x|\theta_0) \} \) denote the rejection region associated with the above rule. Consider any other rule with rejection region \( \mathcal{R}^* \) satisfying \( P(X \in \mathcal{R}^*|\theta = \theta_0) \leq \alpha = P(X \in \mathcal{R}|\theta = \theta_0) \) which also implies:

\[
P(X \in \mathcal{R}^c \cap \mathcal{R}^*|\theta = \theta_0) \leq P(X \in \mathcal{R}^c \cap (\mathcal{R}^*)^c|\theta = \theta_0).
\]

(1)

We need to show \( P(X \in \mathcal{R}^*|\theta = \theta_1) \leq P(X \in \mathcal{R}|\theta = \theta_1) \) which is equivalent to showing

\[
P(X \in \mathcal{R}^c \cap \mathcal{R}^*|\theta = \theta_1) \leq P(X \in \mathcal{R}^c \cap (\mathcal{R}^*)^c|\theta = \theta_1).
\]

(2)

Now, it turns out that for any subset \( A \subset \mathcal{S} \),

\[
P(X \in \mathcal{R}^c \cap A|\theta = \theta_1) = \int_{\mathcal{R}^c \cap A} f(x|\theta_1) dx
\]

\[
\leq c \int_{\mathcal{R}^c \cap A} f(x|\theta_0) dx
\]

\[
= cP(X \in \mathcal{R}^c \cap A|\theta = \theta_0)
\]

and similarly,

\[
P(X \in \mathcal{R} \cap A|\theta = \theta_1) \geq cP(X \in \mathcal{R} \cap A|\theta = \theta_0)
\]

Therefore, the left hand side of (2) is smaller than or equal to \( cP(X \in \mathcal{R}^c \cap \mathcal{R}^*|\theta = \theta_0) \) and the right hand side is greater than or equal to \( cP(X \in \mathcal{R} \cap (\mathcal{R}^*)^c|\theta = \theta_0) \). This establishes (2) because of (1).

\( \Box \)

While the N-P lemma is somewhat restricted in its use of a simple \( H_0 \) and a simple \( H_1 \), it is a profound and influential result. It embodies the “pursuit of optimal rules” on which classical mathematical statistics has thrived for many decades. It also makes the extremely important connection between optimal rules and the likelihood function.

**Example.** Consider again testing \( H_0 : \mu = \mu_0 \) vs. \( H_1 : \mu = \mu_1 \) based on \( X \sim \text{Normal}(\mu,1) \) where \( \mu_0 < \mu_1 \) are two given numbers. Then,

\[
\Lambda(X) = \frac{\exp\left\{-\frac{(X-\mu_1)^2}{2}\right\}}{\exp\left\{-\frac{(X-\mu_0)^2}{2}\right\}} = \exp\left\{(\mu_1-\mu_0)(X-\frac{\mu_0+\mu_1}{2})\right\}
\]

which is an increasing function of \( X \) because \( \mu_1 > \mu_0 \). And hence for every \( c \), the test rule “reject \( H_0 \) if \( X \geq c \)” is UMP at level \( \alpha = P(X \geq c|\mu_0) = 1 - \Phi(c - \mu_0) \). In other words, the UMP test at level \( \alpha \) is given by “reject \( H_0 \) if \( X \geq \mu_0 + c(\alpha) \)” where \( c(\alpha) = \Phi^{-1}(1 - \alpha) \).

Below we present a result that extends N-P lemma to a more interesting situation of a composite \( H_1 \).
Corollary. Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \in \Theta_1$ based on $X \sim f(x|\theta)$. Suppose there is a statistic $T(X)$ such that for every $\theta_1 \in \Theta_1$,

$$\Lambda_{\theta_1}(X) = \frac{f(X|\theta_1)}{f(X|\theta_0)}$$

is an increasing function of $T(X)$. Then for any $c > 0$, the testing rule “reject $H_0$ if $T(X) \geq c$” is UMP at level $\alpha = P(T(X) \geq c|\theta = \theta_0)$. Furthermore, using the test statistic $T(X)$ is equivalent to using the maximum likelihood test statistic $\Lambda(X) = \max_{\theta_1 \in \Theta_1} \Lambda_{\theta_1}(X)$.

Proof. The first assertion is trivial because for every $\theta_1 \in \Theta_1$, the given rule is UMP level $\alpha$ for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. To prove the second assertion, notice that for every $\theta_1 \in \Theta_1$, $\Lambda_{\theta_1}(X) = g_{\theta_1}(T(X))$ for some increasing function $g_{\theta_1}$ and hence $\Lambda(X) = g(T(X))$ where $g(t) := \max_{\theta_1 \in \Theta_1} g_{\theta_1}(t)$, which also is an increasing function. Therefore the rule “reject $H_0$ if $T(X) \geq c$” is equivalent to the rule “reject $H_0$ if $\Lambda(X) \geq k$” where $k = g(c)$. \(\square\)

Example. Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$ based on $X \sim \text{Normal}(\mu, 1)$ where $\mu_0$ is a given number. Then the assumption of the Corollary is satisfied with $T(X) = X$ [see calculations in the earlier example]. And so again the UMP test at level $\alpha$ is given by “reject $H_0$ if $X \geq \mu_0 + c(\alpha)$” with $c(\alpha) = \Phi^{-1}(1 - \alpha)$.

However, the concept of UMP is somewhat limited in its utility because in many realistic situations a UMP does not exist. Consider the following example.

Example. Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ based on $X \sim \text{Normal}(\mu, 1)$. Consider the two rules: “reject $H_0$ if $X \geq \mu_0 + c(\alpha)$” and “reject $H_0$ if $X \leq \mu_0 - c(\alpha)$”; either of which has size $\alpha$ and any other rule with size $\alpha$ or smaller has smaller power than one of these two rules at every $\mu \neq \mu_0$. So no other rule could be UMP. And neither of these two rules is UMP because the first has smaller power than the other at any $\mu < \mu_0$ and the second has smaller power than the first at any $\mu > \mu_0$. Although no UMP rule exists, the maximum likelihood testing rules “reject $H_0$ if $LR(X) \geq k$” offer very reasonable testing rules for this example. Note that a size $\alpha$ ML testing rule in this case must be given by “reject $H_0$ if $|X - \mu_0| \geq c(\alpha/2)$”. Figure 2 shows the two power functions of this rule along with the above two rules for $\mu_0 = 0$ and $\alpha = 5\%$ (for which $c(\alpha) = 1.64$ and $c(\alpha/2) = 1.96$). In fact, it turns out that the ML rule is UMP within the sub-class of “unbiased” test rules which satisfy: $\min_{\mu \neq \mu_0} \text{pow}(\mu) \geq \text{pow}(\mu_0)$.

Due to its limited applicability, we will not pursue the notion of UMP any further, and instead compare testing and other inference rules by means of other metrics, such as consistency and asymptotic efficiency. However, to wrap up the discussion of UMP tests, here is one last result that further extends the Corollary presented above to “one sided tests” on scalar parameters.

**Definition 2 (Monotone Likelihood Ratio).** A statistical model $X \sim f(x|\theta)$ with a scalar parameter $\theta \in \Theta$ is said to possess the monotone likelihood ratio (MLR) property in a statistic $T(X)$ if for every $\theta_1 < \theta_2$, $f(X|\theta_2)/f(X|\theta_1)$ is an increasing function of $T(X)$.

**Theorem 2.** Consider testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ based on $X \sim f(x|\theta)$, where $\theta \in \Theta$ is a scalar parameter. Suppose the statistical model possesses MLR in some statistic $T(X)$. Then for every $c$, the rule “reject $H_0$ if $T(X) \geq c$” is UMP at level $\alpha = P(T(X) \geq c|\theta = \theta_0)$ and is indeed an ML test.
Figure 2: No UMP test exists for testing $H_0: \mu = 0$ vs. $H_1: \mu \neq 0$ based on $X \sim \text{Normal}(\mu, 1)$, but the ML test makes a very reasonable choice.

Proof. Because of the Corollary we proved earlier, the only non-trivial task toward showing “reject $H_0$ if $T(X) \geq c$” is UMP at level $\alpha$ is to show the test indeed has size $\alpha$, i.e., to show that $P(T(X) \geq c|\theta = \theta^*) \leq P(T(X) \geq c|\theta = \theta_0)$ for every $\theta^* < \theta_0$. We show this by proving a stronger result that for any $\theta_1 < \theta_2$, $P(T(X) \geq c|\theta = \theta_1) \leq P(T(X) \geq c|\theta = \theta_2)$.

Let $g(t)$ be any non-decreasing function. Then for any $\theta_1 < \theta_2$ we must have $E(g(T)|\theta = \theta_2) \geq E(g(T)|\theta = \theta_1)$. To see this, let $A = \{x : f(x|\theta_2) \geq f(x|\theta_1)\}$ and define $a = \inf_{x \in A} g(T(x))$, $b = \sup_{x \in A^c} g(T(x))$. By MLR, $a \geq b$ and therefore,

$$E(g(T)|\theta_2) - E(g(T)|\theta_1) = \int g(T(x))\{f(x|\theta_2) - f(x|\theta_1)\}dx$$

$$= \int_A g(T(x))\{f(x|\theta_2) - f(x|\theta_1)\}dx + \int_{A^c} g(T(x))\{f(x|\theta_2) - f(x|\theta_1)\}dx$$

$$\geq a \int_A \{f(x|\theta_2) - f(x|\theta_1)\}dx + b \int_{A^c} \{f(x|\theta_2) - f(x|\theta_1)\}dx$$

$$= (a - b) \int_A \{f(x|\theta_2) - f(x|\theta_1)\}dx$$

where the last equality follows because $\int_A \{f(x|\theta_2) - f(x|\theta_1)\}dx + \int_{A^c} \{f(x|\theta_2) - f(x|\theta_1)\}dx = \int f(x|\theta_2)dx - \int f(x|\theta_1)dx = 0$.

Now apply this result to the non-decreasing function $g(t) = I(t \geq c)$: the indicator function whose value is 0 when $t < c$ and 1 when $t \geq c$.

Example. Consider testing $H_0: p \leq 0.5$ vs. $H_1: p > 0.5$ based on $X \sim \text{Binomial}(n, p)$. For any $p_1 < p_2$,

$$\frac{f(x|p_2)}{f(x|p_1)} = \frac{p_2^n(1 - p_2)^{n-x}}{p_1^n(1 - p_1)^{n-x}} = \frac{(1 - p_2)^n}{(1 - p_1)^n} \times \left(\frac{p_2}{p_1}\right)^x$$
where \( o_j = \frac{p_j}{1-p_j}, j = 1,2 \) are the odds. But \( p_2 > p_1 \) implies \( o_2 > o_1 \), and so the model has MLR in the statistic \( X \). And hence the UMP test at level \( \alpha \) is given by “reject \( H_0 \) if \( X \geq c(\alpha) \)” where \( c(\alpha) \) satisfies \( P(X \geq c(\alpha)|p = 0.5) = \alpha \). For \( n \) large, the distribution of \( X \) under \( p = 0.5 \) is well approximated by Normal\((n/2,n/4)\) and hence we can approximate \( c(\alpha) \approx n/2 + (\sqrt{n}/2) \cdot \Phi^{-1}(1-\alpha) \).

### 4 Exponential family models

Possessing MLR property is more of an exception than norm. A simple example is the Cauchy model: \( X \sim \text{Cauchy}(\theta, 1), \theta \in (-\infty, \infty) \) with pdf \( f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)} \) which is not MLR in \( X \). However a fair number of useful models exhibit MLR. These are typically given by an exponential family of distributions.

**Definition 3 (Exponential family).** A family of pdfs/pmfs \( f(x|\theta), \theta \in \Theta \) on \( S \) is called an exponential family if there exist functions \( h: S \rightarrow \mathbb{R}_+, T: S \rightarrow \mathbb{R}^q, \eta: \Theta \rightarrow \mathbb{R}^q \) and \( B: \Theta \rightarrow \mathbb{R} \) such that

\[
f(x|\theta) = h(x) \exp\{\eta(\theta)^T T(x) - B(\theta)\}, \quad \text{for every } x \in S, \theta \in \Theta.
\]

A number of useful statistical models are based on exponential families, such as binomial, normal, Poisson, exponential, gamma, exponential etc [see http://en.wikipedia.org/wiki/Exponential_family]. Note that if \( X = (X_1, \ldots, X_n) \) is modeled as \( X_i \overset{iid}{\sim} g(x_i|\theta) \) with \( g(x_i|\theta) = h(x_i) \exp\{\eta(\theta)^T T(x_i) - B(\theta)\} \) then the joint pdf/pmf of \( X \) equals:

\[
f(x|\theta) = h^*(x) \exp\{\eta(\theta)^T T^*(x) - B^*(\theta)\}
\]

where \( h^*(x) = h(x_1) \times \cdots \times h(x_n), T^*(x) = T(x_1) + \cdots + T(x_n) \) and \( B^*(\theta) = nB(\theta) \). For exponential family models, MLRs and UMPs arise naturally.

**Lemma 3.** A scalar exponential family model: \( X \sim f(x|\theta) = h(x) \exp\{\eta(\theta)^T T(x) - B(\theta)\}, \theta \in \Theta \subset \mathbb{R} \) with a monotone increasing map \( \theta \mapsto \eta(\theta) \) possesses MLR in \( T(X) \).

**Proof.** This follows easily because \( f(x|\theta_2)/f(x|\theta_1) = \exp\{[\eta(\theta_2) - \eta(\theta_1)]T(x) - \{B(\theta_2) - B(\theta_1)\}\} \) which is increasing in \( T(x) \) if \( \theta_2 > \theta_1 \). \( \Box \)