3.1 Agenda: Global Testing

1. Simes Test
2. Tests based on Empirical CDF’s
   (a) Kolmogorov-Smirnov Test
   (b) Anderson-Darling Test
   (c) Tukey’s Second-Level Significance Testing
3. Sparse Mixtures

3.2 Simes Test

As in the previous lectures, consider \( n \) hypotheses \( H_{0,i} \) and \( p \)-values \( p_i \). We are interested in testing the global null \( H_0 = \cap_{i=1,...,n} H_{0,i} \), where under \( H_{0,i} \), \( p_i \sim \text{Unif}(0,1) \).

The Simes test, which was simultaneously introduced by Simes [10] and Eklund [5], is a modification of the Bonferroni procedure based on ordered \( p \)-values \( p^{(1)} \leq p^{(2)} \leq \ldots \leq p^{(n)} \). We will see in future lectures that the Simes test is connected to the Benjamini-Hochberg procedure [2].

The Simes statistic is given by

\[
T_n = \min_i \{ p^{(i)} \frac{n}{i} \},
\]

where \( n/i \) is an adjustment factor, \( i = 1, ..., n \). Note that this adjustment factor is equal to one for the largest \( p \)-value, but equal to \( n \) for the smallest \( p \)-value.

**Theorem 1.** Under \( H_0 \) and independence of the \( p_i \)'s, \( T_n \sim \text{Unif}(0,1) \). Thus, the Simes test rejects \( H_0 \) if \( T_n \leq \alpha [10] \).

As we will see next week, the Simes test also has size at most \( \alpha \) under a sort of positive dependence [9]. This follows from the connection of the Simes test to the Benjamini-Hochberg procedure. However, in this lecture, we will focus on the independent case.
Note that

\[ T_n = \min_i \left\{ p(i) \frac{n}{i} \right\} \iff \exists i : p(i) \leq \frac{i}{n}. \]

Hence, equivalently, we could reject \( H_0 \) if \( \exists i : p(i) \leq \frac{i}{n} \).

The sketch in figure 3.1 illustrates this idea. In the left panel, no \( p \)-value falls below the critical line, hence \( H_0 \) would not be rejected. In the right panel, there exists a \( p \)-value which falls below the critical line and therefore we would reject \( H_0 \).

Figure 3.1. Sketch of an example where the Simes test would not reject \( H_0 \) (left panel) and where it would reject (right panel). \( p \)-values are displayed as grey dots, the critical line \( \frac{i}{n} \) is shown in grey. The \( p \)-value below the critical line in the right panel is shown as a red dot.

The Simes procedure is strictly less conservative than Bonferroni, which rejects when \( p(1) \leq \frac{\alpha}{n} \). This implies that the rejection region of the Simes test includes the rejection region of Bonferroni (and more). For \( n = 2 \), we can visualize the rejection region as in figure 3.2 below.

Figure 3.2. Rejection Regions for Simes and Bonferroni Procedures for \( n = 2 \)

In this example, the Simes test rejects if \( p(1) \leq \frac{\alpha}{2} \) or \( p(2) \leq \alpha \). Thus, the rejection region of the Simes test is larger than the rejection region of Bonferroni. The picture also confirms theorem 1 for \( n = 2 \) (this can be seen by adding up the areas of interest): The Simes statistic actually is \( \text{Unif}(0, 1) \) when \( n = 2 \).
However, note that with respect to power, the Simes procedure has similar properties to Bonferroni: It is powerful for a few strong effects and has moderate power for many mild effects.

Next, we want to prove theorem 1.

Proof. We prove theorem 1 by induction. Following Simes [10], the theorem is clearly true for \( n = 1 \). The above example also shows that it holds for \( n = 2 \), see figure 3.2. Assume that the theorem holds for \( n - 1 \). We then want to show that the theorem holds for \( n \). Under the induction hypothesis, \( T_{n-1} \sim \text{Unif}(0, 1) \).

Recall that the \( p \)-values are ordered \( p(1) \leq p(2) \leq \ldots \leq p(n) \). The density of \( p(n) \) is \( f(t) = nt^{n-1} \) for \( t \in [0, 1] \).

\[
P(T_n \leq \alpha) = P(\min_i \{p(i) \frac{n}{i}\} \leq \alpha)
= P(p(n) \leq \alpha) + P(T_n \leq \alpha, p(n) > \alpha)
= P(p(n) \leq \alpha) + P(p(n) > \alpha, \min_{1 \leq i \leq n-1} p(i) \frac{n}{i} \leq \alpha)
\]

We will consider each term separately. First,

\[
P(p(n) \leq \alpha) = \int_0^\alpha nt^{n-1} dt
= \frac{\alpha^n}{n}
\]

Next, consider the second term. Conditional on \( p(n) = t \), the other \( p \)-values are independently \( \text{Unif}(0, t) \). Moreover,

\[
\min_{1 \leq i \leq n-1} p(i) \frac{n}{i} \leq \alpha \iff \min_{1 \leq i \leq n-1} \frac{p(i) n - 1}{i} \leq \frac{\alpha n - 1}{n}.
\]

Hence, we can apply the induction hypothesis:

\[
P(p(n) > \alpha, \min_{1 \leq i \leq n-1} p(i) \frac{n}{i} \leq \alpha) = \int_\alpha^1 \frac{\alpha}{t} \int_0^\alpha P(T_n \leq \alpha|p(n) = t)f(t)dt
= \int_\alpha^1 \frac{\alpha}{t} \int_0^\alpha nt^{n-1} dt
= \int_\alpha^1 \frac{\alpha}{t} \alpha n - 1 \frac{1}{n} t^{n-1} dt
= \alpha \int_\alpha^1 (n-1) t^{n-2} dt
= \alpha(1 - \alpha^{n-1})
= \alpha - \alpha^n
\]

3-3
Putting the results together, we obtain

\[
P(T_n \leq \alpha) = P(\min_i \{p(i)n\} \leq \alpha)
= P(p(n) \leq \alpha) + P(p(n) > \alpha, \min_{1 \leq i \leq n-1} \{p(i)n\} \leq \alpha)
= \alpha^n + \alpha - \alpha^n
= \alpha,
\]

which completes the proof.

\[\square\]

3.3 Tests based on Empirical CDF’s

**Definition 1.** The empirical CDF of \(p_1, ..., p_n\) is

\[
\hat{F}_n(t) = \frac{1}{n} \#\{i : p_i \leq t\}.
\]

Intuitively, the empirical CDF is the fraction of elements that fall below or are equal to some threshold \(t\). Under the global null hypothesis \(H_0\), \(1\{p_i \leq t\} \sim Ber(t)\) and therefore \(E(\hat{F}_n(t)) = t\). This implies that under independence \(n\hat{F}_n(t) \sim Bin(n, t)\). The idea is to measure the differences between what we would expect under the null hypothesis and what we actually observe. Therefore, we would reject the null hypothesis if the difference between \(\hat{F}_n(t)\) and \(t\) is large. Building onto this intuition, we now consider three different tests based on the empirical CDF: The Kolmogorov-Smirnov Test (3.3.1), the Anderson-Darling Test (3.3.2) and Tukey’s Second-Level Significance Test (3.3.3).

3.3.1 Kolmogorov-Smirnov Test

**Definition 2.** The Kolmogorov-Smirnov (KS) test statistic is defined as

\[
KS = \sup_t |\hat{F}_n(t) - t|.
\]

The null hypothesis is rejected if the KS test statistic exceeds a certain threshold. Note that we are usually interested in rejecting if \(KS^+ = \sup_t (\hat{F}_n(t) - t)\) exceeds a certain threshold. The reason is that we would like to reject for small \(p\)-values, i.e. large values of \(\hat{F}_n(t) - t\).

To find the correct threshold, we would need to know the distribution under the global null hypothesis, which is not known in closed form. In real life, one might compute simulations or rely on asymptotic calculations.

A useful inequality was developed by Massart [8], who showed that the constant \(C\) in the Dvoretzky, Kiefer and Wolfowitz inequality [4] can be taken to be equal to one.
Theorem 2. Massart’s inequality [8]: Under $H_0$ and independence,

$$\mathbb{P}(KS^+ \geq u) \leq e^{-2nu^2}$$

for $u \geq \sqrt{\frac{\log 2}{2n}}$.

Note that $u \geq \sqrt{\frac{\log 2}{2n}}$ will usually hold for us. Thus, the tail of the Kolmogorov-Smirnov statistic is sub-Gaussian and therefore decays fast.

3.3.2 Anderson-Darling Test

Definition 3. Consider the test-statistic defined by

$$A^2 = n \int_0^1 (\hat{F}_n(t) - t)^2 \omega(t)dt,$$

where $\omega(t) \geq 0$ is a weight function.

If $\omega(t) = 1$, the above statistic is called the Cramer-von Mises statistic.

The sketch in figure 3.3 illustrates the main idea: Instead of looking at the value of $t$ for which the difference between the empirical CDF and $t$ is largest, as in the KS test, the evidence across all values of $t$ is considered (shaded area).

Anderson and Darling [1] consider $\omega(t) = [t(1-t)]^{-1}$. Thus,

$$A^2 = n \int_0^1 (\hat{F}_n(t) - t)^2 \frac{t}{t(t-1)}dt.$$

Under the global null, $n\hat{F}_n(t) \sim \text{Bin}(n,t)$. In particular, $Var(\hat{F}_n(t)) \propto t(1-t)$. Hence, $\frac{(\hat{F}_n(t)-t)^2}{t(t-1)}$ intuitively is similar to a (squared) z-score: The (squared) z-score is “averaged” over $t$ and the null hypothesis is rejected if this statistic is large. Since $\omega(t) = [t(1-t)]^{-1}$, the Anderson-Darling statistic puts more weight on small/large $p$-values than the Cramer-von Mises statistic.
Recall that Fisher’s test statistic is $T_{\text{Fisher}} = -2 \sum_{i=1}^{n} \log(p_i)$ and the Pearson test statistic is $T_{\text{Pearson}} = -2 \sum_{i=1}^{n} \log(1-p_i)$. In order to connect the Anderson-Darling test to the test statistics we have seen in the previous lectures, note the following useful relationship:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[ \log(p_{(i)}) + \log(1-p_{(n+1-i)}) \right]$$

Thus, the Anderson-Darling test can be seen as a combination of Fisher’s test and the Pearson test. While in Fisher’s statistic the logged $p$-values receive the same weight regardless of their size, the Anderson-Darling statistic re-weights them depending on their rank. Recall that the Fisher’s test was sensitive to small values of the $p$-values. By re-weighting, the Anderson-Darling test alleviates this problem. In summary, the Anderson-Darling test gives more weight to $p$-values that are in the bulk compared to Fisher’s statistic.

### 3.3.3 Tukey’s Second-Level Significance Testing (Higher-criticism statistic)

As we have seen earlier (3.3.1), the Kolmogorov-Smirnov test looks for the maximum distance between the empirical CDF and its expected value under the null-hypothesis, while the Anderson-Darling test integrates the differences instead. We now want to combine these approaches.

Again, suppose that we were to test $n$ hypotheses at level $\alpha$. We would expect $n\alpha$ tests to be significant, while the observed number of significant tests would be $\hat{F}_n(\alpha)$ and the standard deviation would be given by $\sqrt{n\alpha(1-\alpha)}$. Thus, as previously, we construct a $z$-score and the overall significance at level $\alpha$ would be

$$z = \frac{(\# \text{ significant tests at level } \alpha) - \text{ expected}}{\text{SD}} = \frac{n\hat{F}_n(\alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)/n}}.$$ 

According to Donoho and Jin [3], Tukey [11] proposed in his class notes to use a second-level significance testing. Compared to the Anderson-Darling statistic, the higher-criticism statistic uses the maximum value instead of a (squared) average.

**Definition 4.** The higher-criticism statistic [11] is:

$$HC^*_n = \max_{0<\alpha<\alpha_0} \frac{\hat{F}_n(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)/n}}.$$ 

Donoho and Jin [3], based on Jin [7] and Ingster [6], provide a theoretical analysis of Tukey’s higher-criticism statistic. The higher-criticism statistic “scans” across the significance levels for departures from $H_0$. Hence, a large value of $HC^*_n$ indicates significance of an overall body of tests.

The question now is whether we can actually do better than Bonferroni. To understand the power of the higher-criticism statistic, we will study sparse mixtures next.
3.4 Sparse Mixtures

Donoho and Jin [3], based on Jin [7] and Ingster [6], consider $n$ tests of $H_{0,i}$ vs. $H_{1,i}$ with independent test statistics $X_i$.

The original model is:

\[ H_{0,i} : X_i \sim \mathcal{N}(0,1) \]
\[ H_{1,i} : X_i \sim \mathcal{N}(\mu_i,1), \ \mu_i > 0 \]

We are interested in possibilities within $H_1$ with a small fraction of non-null hypotheses. Rather than directly assuming that there is some amount of nonzero means under $H_1$, we assume that our samples follow a mixture of $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$, with $\mu > 0$ fixed and parameter $\epsilon$.

This simple model with equal means can be written as:

\[ H_0 : X_i \text{i.i.d.} \sim \mathcal{N}(0,1) \]
\[ H_1 : X_i \text{i.i.d.} \sim (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu,1) \]

The expected number of non-nulls under $H_1$ is $n\epsilon$. If $\epsilon = 1/n$, then the above would become the “needle in a haystack” problem: On average, there would be one coordinate with $\mu$ nonzero.

If $\epsilon$ and $\mu$ were known, then the optimal test would be the likelihood ratio test. The likelihood ratio under the sparse mixture model is

\[ L = \prod_{i=1}^{n} [(1-\epsilon) + \epsilon e^{\mu X_i - \mu^2/2}] \]

The asymptotic analysis of Ingster [6] and Jin [7] specifies the dependency of $\epsilon$ and $\mu$ on $n$ as follows:

\[ \epsilon_n = n^{-\beta} \]
\[ \mu_n = \sqrt{2r \log n} \]

\[ 1 < \beta < 1 \]
\[ 0 < r < 1 \]

The parameter $\beta$ controls the sparsity of the alternative, while $r$ parameterizes the mean shift. If $\beta$ were large, then our problem would be very sparse, while if $\beta$ were small, it would be mildly sparse. If $r = 1$, then we get the detection threshold we have seen for Bonferroni. Hence, the “needle in a haystack” problem corresponds to $\beta = 1$ and $r = 1$.

Ingster [6] and Jin [7] find that there is a threshold curve for $r$ of the form
\[ \rho^*(\beta) = \begin{cases} 
\beta - 1/2 & \frac{1}{2} < \beta \leq \frac{3}{4} \\
(1 - \sqrt{1 - \beta})^2 & \frac{3}{4} \leq \beta \leq 1 
\end{cases} \]

such that

1. If \( r > \rho^*(\beta) \) we can adjust the NP test to achieve
   \[ \mathbb{P}_0(\text{Type I Error}) + \mathbb{P}_1(\text{Type II Error}) \to 0 \]

2. If \( r < \rho^*(\beta) \) then for any test
   \[ \liminf \mathbb{P}_0(\text{Type I Error}) + \mathbb{P}_1(\text{Type II Error}) \geq 1. \]

Unfortunately, we generally cannot use the NP test, since we do not know \( \epsilon \) or \( \mu \). However, Donoho and Jin [3], based on Ingster [6], Jin [7] show that Tukey’s higher-criticism statistic, which does not require knowledge of \( \epsilon \) or \( \mu \), asymptotically achieves the optimal detection threshold, with

\[ p_i = \Phi(X_i) = \mathbb{P}(N(0,1) > X_i) \]

\[ HC^*_n = \max_{\alpha \leq \alpha_0} \frac{\sqrt{n}(F_n(\alpha) - \alpha)}{\sqrt{\alpha(1 - \alpha)}}. \]

To better understand the results above, figure 3.4 visualizes the detection thresholds for NP and Bonferroni.

**Figure 3.4.** Detection Thresholds for NP and Bonferroni Tests

If the amplitude of the signal is above the solid black curve (achievable with NP), then the NP test has full power, that is, we asymptotically separate. However, if it is below the curve, we asymptotically merge (every test is no better than tossing a coin).
The dashed black curve in figure 3.4 shows the detection threshold for Bonferroni. Bonferroni’s method achieves the optimal threshold for \( \beta \in [\frac{3}{4}, 1] \) (sparsest setting), but has a suboptimal threshold if \( \beta \in [\frac{1}{2}, \frac{3}{4}] \) (less sparsity). This is also visualized in the figure, since the dashed Bonferroni curve and the solid NP curve align for \( \beta \geq \frac{3}{4} \), but separate below. Hence, in the area between the Bonferroni and the NP curve, the NP test has full power, while Bonferroni is no better than a coin toss.

Bonferroni’s threshold for \( \frac{1}{2} \leq \beta \leq 1 \) is

\[
\rho_{\text{Bon}}(\beta) = (1 - \sqrt{1 - \beta})^2.
\]

Bonferroni is powerless if \( r < \rho_{\text{Bon}} \).

The correct detection happens if the maximum of non nulls is greater than that of nulls, i.e. roughly

\[
\max_{\text{non null}} X_i \simeq \sqrt{2r \log n + \sqrt{2} \log n^{1-\beta}} > \sqrt{2 \log n}
\]

\[
\Leftrightarrow \sqrt{r + 1 - \beta} > 1
\]

\[
\Leftrightarrow r > (1 - \sqrt{1 - \beta})^2 = \rho_{\text{Bon}}(\beta).
\]

The higher-criticism test rejects when \( HC^*_n \) is large, i.e. when the \( p \)-values tend to be a bit too small. Next, we will consider “how small” the \( p \)-value should be, which will be discussed in more detail in the next lecture.

For now, consider the empirical process

\[
W_n(t) = \frac{\sqrt{n}(F_n(t) - t)}{\sqrt{t(1-t)}},
\]

where \( W_n(t) \) converges in distribution to \( N(0, 1) \) for each \( t \). However, note that \( W_n(t) \) is heavy tailed near 0, as shown in figure 3.5.

Empirical process theory tells us that:

1. \( \{\sqrt{n}(F_n(t) - t)\}_{0 \leq t \leq 1} \) converges in distribution to a Brownian bridge

2. \( \max_{1/n \leq t \leq \alpha_0} W_n(t)/\sqrt{2 \log \log n} \xrightarrow{p} 1 \)

This suggests the threshold \( \sqrt{2 \log \log n} \) for the HC statistic.

**Theorem 3.** (Donoho and Jin [3]): If we reject when \( HC^*_n \geq \sqrt{(1 + \epsilon)2 \log \log n} \), then for any alternative \( H_1 \) in which \( r > \rho^*(\beta) \),

\[
\mathbb{P}_0(\text{Type I Error}) + \mathbb{P}_1(\text{Type II Error}) \rightarrow 0
\]

3-9
One problem of the higher-criticism statistic in practice is that the value of the threshold is quite dominated by what happens at small values of $t$. We will discuss shortcomings of the higher-criticism statistic in more detail during the next lecture.
References


