From Hypothesis Tests to Confidence Regions

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

1. p-values
2. Confidence Intervals
3. The Pivoting Method
4. Extension to Confidence Regions
5. Inverting Hypothesis Tests
6. Multiple testing

p-values

Beyond Neyman-Pearson?

So far restricted to Neyman-Pearson Framework:

1. Fix a significance level α for the test
2. Consider rules δ respecting this significance level
   $\hookrightarrow$ We choose one of those rules, δ^*, based on power considerations
3. We reject at level α if $\delta^*(\mathbf{x}) = 1$.

Useful for attempting to determine optimal test statistics

What if we already have a given form of test statistic in mind? (e.g. LRT)

$\hookrightarrow$ A different perspective on testing (used more in practice) says:

> Rather then consider a family of test functions respecting level α...
> ... consider family of test functions indexed by α

1. Fix a family $\{\delta_\alpha\}_{\alpha\in(0,1)}$ of decision rules, with δ_α having level α
   $\hookrightarrow$ for a given $\mathbf{x}$ some of these rules reject the null, while others do not
2. Which is the smallest α for which H_0 is rejected given $\mathbf{x}$?

Observed Significance Level

Definition (p–Value)

Let $\{\delta_\alpha\}_{\alpha\in(0,1)}$ be a family of test functions satisfying

$$\alpha_1 < \alpha_2 \implies \{\mathbf{x} \in \mathcal{X} : \delta_{\alpha_1}(\mathbf{x}) = 1\} \subseteq \{\mathbf{x} \in \mathcal{X} : \delta_{\alpha_2}(\mathbf{x}) = 1\}.$$

The p–value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\mathbf{x}) = \inf\{\alpha : \delta_\alpha(\mathbf{x}) = 1\}$$

$\hookrightarrow$ The p–value is the smallest value of α for which the null would be rejected at level α, given $\mathbf{X} = \mathbf{x}$.
Most usual setup:

- Have a single test statistic T
- Construct family $\delta_\alpha(\mathbf{x}) = \mathbf{1}\{T(\mathbf{x}) > k_\alpha\}$
- If $\mathbb{P}_{H_0}[T \leq t] = G(t)$ then $p(\mathbf{x}) = \mathbb{P}_{H_0}[T(\mathbf{X}) \geq T(\mathbf{x})] = 1 - G(T(\mathbf{x}))$

Observed Significance Level

Notice: contrary to NP-framework did not make explicit decision!

- We simply reported a p–value
- The p–value is used as a measure of evidence against H_0
  - $\hookrightarrow$ Small p–value provides evidence against H_0
  - $\hookrightarrow$ Large p–value provides no evidence against H_0
- How small does "small" mean?
  - $\hookrightarrow$ Depends on the specific problem...

Intuition:

- Recall that extreme values of test statistics are those that are "inconsistent" with null (NP-framework)
- p–value is probability of observing a value of the test statistic as extreme as or more extreme than the one we observed, under the null
- If this probability is small, the we have witnessed something quite unusual under the null hypothesis
- Gives evidence against the null hypothesis

Example (Normal Mean)

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0$$

Likelihood ratio test: reject when T^2 large, $T = \sqrt{n}\bar{X}/S \overset{H_0}{\sim} t_{n-1}$.

Since $T^2 \overset{H_0}{\sim} F_{1,n-1}$, p–value is

$$p(\mathbf{x}) = \mathbb{P}_{H_0}[T^2(\mathbf{X}) \geq T(\mathbf{x})] = 1 - G_{F_{1,n-1}}(T^2(\mathbf{x}))$$

Consider two samples (datasets),

$$\mathbf{x} = (0.66, 0.28, -0.99, 0.007, -0.29, -1.88, -1.24, 0.94, 0.53, -1.2)$$

$$\mathbf{y} = (1.4, 0.48, 2.86, 1.02, -1.38, 1.42, 2.11, 2.77, 1.02, 1.87)$$

Obtain $p(\mathbf{x}) = 0.32$ while $p(\mathbf{y}) = 0.006$.

Significance VS Decision

- Reporting a p–value does not necessarily mean making a decision
- A small p–value can simply reflect our "confidence" in rejecting a null
  $\hookrightarrow$ reflects how statistically significant the alternative statement is

Recall example: Statisticians working for Obama gather iid sample $\mathbf{X}$ from Ohio with $X_i = \mathbf{1}\{\text{vote Obama}\}$. Obama team want to test

$$\begin{cases} H_0: & \text{Romney wins Ohio} \\ H_1: & \text{Obama wins Ohio} \end{cases}$$

- Will statisticians decide for Obama?
- Perhaps better to report p–value to him and let him decide...

What if statisticians working for newspaper, not Obama?

- Something easier to interpret than test/p–value?

Confidence Intervals

A Glance Back at Point Estimation

- Let $X_1, ..., X_n$ be iid random variables with density (frequency) $f(\cdot; \theta)$.
- Problem with point estimation: $\mathbb{P}_\theta[\hat{\theta} = \theta]$ typically small (if not zero)
  - $\hookrightarrow$ always attach an estimator of variability, e.g. standard error
  - $\hookrightarrow$ interpretation?

- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem
  - $\hookrightarrow$ e.g. if observe $\hat{P}[\text{obama wins}] = 0.52$ can actually see what p–value we get when testing $H_0 : P[\text{obama wins}] \geq 1/2$.

- Something more directly interpretable?

Back to our example: What do pollsters do in newspapers?

$\hookrightarrow$ They announce their point estimate (e.g. 0.52)

$\hookrightarrow$ They give upper and lower confidence limits

What are these and how are they interpreted?

Interval Estimation

Simple underlying idea:

- Instead of estimating θ by a single value
- Present a whole range of values for θ that are consistent with the data
  - $\hookrightarrow$ In the sense that they could have produced the data

Definition (Confidence Interval)

Let $\mathbf{X} = (X_1, ..., X_n)$ be random variables with joint distribution depending on $\theta \in \mathbb{R}$ and let $L(\mathbf{X})$ and $U(\mathbf{X})$ be two statistics with $L(\mathbf{X}) < U(\mathbf{X})$ a.s. Then, the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called a $100(1-\alpha)\%$ confidence interval for θ if

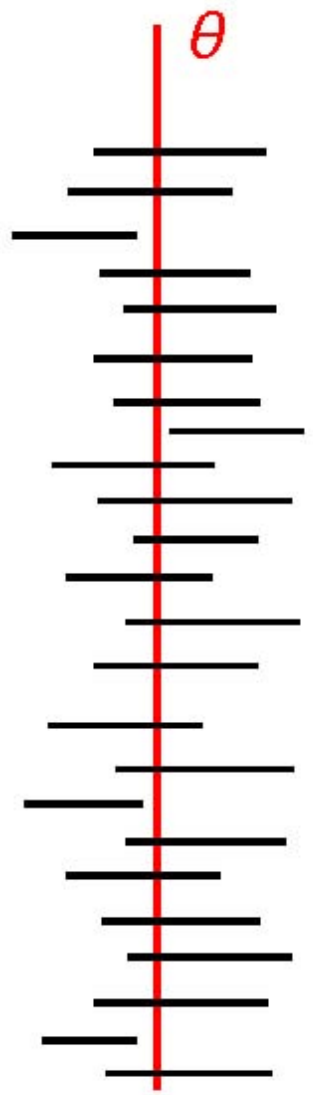
$$\mathbb{P}_\theta[L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})] \geq 1 - \alpha$$

for all $\theta \in \Theta$, with equality for at least one value of θ.

- $1 - \alpha$ is called the coverage probability or confidence level
- Beware of interpretation!

Interval Estimation: Interpretation

- Probability statement is NOT made about θ, which is constant.
- Statement is about interval: probability that the interval contains the true value is at least $1 - \alpha$.
- Given any realization $\mathbf{X} = \mathbf{x}$, the interval $[L(\mathbf{x}), U(\mathbf{x})]$ will either contain or not contain θ.
- Interpretation: if we construct intervals with this method, then we expect that $100(1 - \alpha)\%$ of the time our intervals will engulf the true value.



Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Then $\sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$, so that

$$\mathbb{P}_\mu[-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96] = 0.95$$

and since

$$-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96 \iff \bar{X} - 1.96/\sqrt{n} \leq \mu \leq \bar{X} + 1.96/\sqrt{n}$$

we obviously have

$$\mathbb{P}_\mu \left[\bar{X} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1.96}{\sqrt{n}} \right] = 0.95$$

So that the random interval $[L(\mathbf{X}), U(\mathbf{X})] = \left[\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}} \right]$ is a 95% confidence interval for μ.

Central Limit Theorem: same argument can yield approximate 95% CI when $X_1, ..., X_n$ are iid, $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = 1$, regardless of their distribution.

The Pivoting Method

Pivotal Quantities

What can we learn from previous example?

Definition (Pivot)

A random function $g(\mathbf{X}, \theta)$ is said to be a pivotal quantity (or simply a pivot) if it is a function both of $\mathbf{X}$ and θ whose distribution does not depend on θ.

$\hookrightarrow \sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$ is a pivot in previous example

Why is a pivot useful?

- $\forall\ \alpha \in (0, 1)$ we can find constants $a < b$ independent of θ, such that

$$\mathbb{P}_\theta[a \leq g(\mathbf{X}, \theta) \leq b] = 1 - \alpha \qquad \forall\ \theta \in \Theta$$

- If $g(\mathbf{X}, \theta)$ can be manipulated then the above yields a CI

Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. Recall that MLE $\hat{\theta}$ is $\hat{\theta} = X_{(n)}$, with distribution

$$\mathbb{P}_\theta \left[X_{(n)} \leq x \right] = F_{X_{(n)}}(x) = \left(\frac{x}{\theta} \right)^n \implies \mathbb{P}_\theta \left[\frac{X_{(n)}}{\theta} \leq y \right] = y^n$$

$\rightarrow$ Hence $X_{(n)}/\theta$ is a pivot for θ. Can now choose $a < b$ such that

$$\mathbb{P}_\theta \left[a \leq \frac{X_{(n)}}{\theta} \leq b \right] = 1 - \alpha$$

$\rightarrow$ But there are ∞-many such choices!
$\hookrightarrow$ Idea: choose pair (a, b) that minimizes interval's length!

Solution can be seen to be $a = \alpha^{1/n}$ and $b = 1$, yielding

$$\left[X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}} \right]$$

Comments on Pivotal Quantities

Pivotal method extends to construction of CI for θ_k, when

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_k, ..., \theta_p) \in \mathbb{R}^p$$

and the remaining coordinates are also unknown. $\rightarrow$ Pivotal quantity should now be function $g(\mathbf{X}; \theta_k)$ which

1. Depends on $\mathbf{X}$, θ_k, but no other parameters
2. Has a distribution independent of any of the parameters

$\hookrightarrow$ e.g.: CI for normal mean, when variance unknown

$\rightarrow$ Main difficulties with pivotal method:

- Hard to find exact pivots in general problems
- Exact distributions may be intractable

Resort to asymptotic approximations...

$\hookrightarrow$ Most classic example when have $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$.

Extension to Confidence Regions

Confidence Regions

What about higher dimensional parameters?

Definition (Confidence Region)

Let $\mathbf{X} = (X_1, ..., X_n)$ be random variables with joint distribution depending on $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. A random subset $R(\mathbf{X})$ of Θ depending on $\mathbf{X}$ is called a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ if

$$\mathbb{P}_\theta[R(\mathbf{X}) \ni \boldsymbol{\theta}] \geq 1 - \alpha$$

for all $\boldsymbol{\theta} \in \Theta$, with equality for at least one value of $\boldsymbol{\theta}$.

- No restriction requiring $R(\mathbf{X})$ to be convex or even connected
  - $\hookrightarrow$ So when $p = 1$ get more general notion than CI
- Nevertheless, many notions extend immediately to CR case
  - $\hookrightarrow$ e.g. notion of a pivotal quantity

Pivots for Confidence Regions

Let $g : \mathcal{X} \times \Theta \to \mathbb{R}$ be a function such that dist$[g(\mathbf{X}, \boldsymbol{\theta})]$ independent of $\boldsymbol{\theta}$
$\hookrightarrow$ Since image space is the real line, can find $a < b$ s.t.

$$\mathbb{P}_{\boldsymbol{\theta}}[a \leq g(\mathbf{X}, \boldsymbol{\theta}) \leq b] = 1 - \alpha$$

$$\implies \mathbb{P}_{\boldsymbol{\theta}}[R(\mathbf{X}) \ni \boldsymbol{\theta}] = 1 - \alpha$$

where $R(\mathbf{x}) = \{\boldsymbol{\theta} \in \Theta : g(\mathbf{x}, \boldsymbol{\theta}) \in [a, b])\}$

Notice that region can be "wild" since it is a random fibre of g

Example

Let $\mathbf{X}_1, ..., \mathbf{X}_n \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$. Two unbiased estimators of $\boldsymbol{\mu}$ and Σ are

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^T$$

Example (cont'd)

Consider the random variable

$$g(\{\mathbf{X}\}_{i=1}^n, \boldsymbol{\mu}) := \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \sim F\text{-dist with k and n-k d.f.}$$

A pivot!

$\hookrightarrow$ If f_q is q-quantile of this distribution, then get $100q\%$ CR as

$$R(\{\mathbf{X}\}_{i=1}^n) = \left\{\boldsymbol{\theta} \in \mathbb{R}^n : \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \leq f_q\right\}$$

- An ellipsoid in $\mathbb{R}^n$
- Ellipsoid centred at $\hat{\boldsymbol{\mu}}$
- Principle axis lengths given by eigenvalues of $\hat{\boldsymbol{\Sigma}}^{-1}$
- Orientation given by eigenvectors of $\hat{\boldsymbol{\Sigma}}^{-1}$

Getting Confidence Regions from Confidence Intervals

Visualisation of high-dimensional CR's can be hard

- When these are ellipsoids spectral decomposition helps
- But more generally?

Things especially easy when dealing with rectangles - but they rarely occur!
$\hookrightarrow$ What if we construct a CR as Cartesian product of CI's?

Let $[L_i(\mathbf{X}), U_i(\mathbf{X})]$ be $100q_i\%$ CI's for θ_i, $i = 1, .., p$, and define

$$R(\mathbf{X}) = [L_1(\mathbf{X}), U_1(\mathbf{X})] \times \ldots \times [L_p(\mathbf{X}), U_p(\mathbf{X})]$$

Bonferroni's inequality implies that

$$\mathbb{P}_{\boldsymbol{\theta}}[R(\mathbf{X}) \ni \boldsymbol{\theta}] \geq 1 - \sum_{i=1}^{p} \mathbb{P}[\theta_i \notin [L_i(\mathbf{X}), U_i(\mathbf{X})]] = 1 - \sum_{i=1}^{p}(1 - q_i)$$

$\rightarrow$ So pick q_i such that $\sum_{i=1}^{p}(1 - q_i) = \alpha$ (can be conservative...)

Inverting Hypothesis Tests

Confidence Intervals and Hypothesis Tests

Discussion on CR's $\rightarrow$ no guidance to choosing "good" regions

But: $\exists$ close relationship between CR's and HT's!
$\hookrightarrow$ exploit to transform good testing properties into good CR properties

Suppose $R(\mathbf{X})$ is an exact $100q\%$=$100(1-\alpha)\%$ CR for $\boldsymbol{\theta}$. Consider

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

Define test function:

$$\delta(\mathbf{X}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}_0 \notin R(\mathbf{X}), \\ 0 & \text{if } \boldsymbol{\theta}_0 \in R(\mathbf{X}). \end{cases}$$

Then, $$\mathbb{E}_{\boldsymbol{\theta}_0}[\delta(\mathbf{X})] = 1 - \mathbb{P}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_0 \in R(\mathbf{X})] \leq \alpha$$

Can use a CR to construct test with significance level α!

Confidence Intervals and Hypothesis Tests

Going the other way around, can invert tests to get CR's:

Suppose we have tests at level α for any choice of simple null, $\boldsymbol{\theta}_0 \in \Theta$.
$\hookrightarrow$ Say that $\delta(\mathbf{X}; \boldsymbol{\theta}_0)$ is appropriate test function for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

Define $$R^*(\mathbf{X}) = \{\boldsymbol{\theta}_0 : \delta(\mathbf{X}; \boldsymbol{\theta}_0) = 0\}$$

Coverage probability of $R^*(\mathbf{X})$ is

$$\mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \in R^*(\mathbf{X})] = \mathbb{P}_{\boldsymbol{\theta}}[\delta(\mathbf{X}; \boldsymbol{\theta}) = 0] \geq 1 - \alpha$$

Obtain a $100(1-\alpha)\%$ confidence region by choosing all the $\boldsymbol{\theta}$ for which the null would not be rejected given our data $\mathbf{X}$.

$\hookrightarrow$ If test inverted is powerful, then get "small" region for given $1 - \alpha$.

Summary

p-values provides an alternative framework for hypothesis testing:

- Strong point: more nuanced judgement on H_0.
- Weakness: users usually forget about power.
- Key point: in the right hands, p-values are innocuous.
  In the wrong hands though ...

Confidence intervals provide a richer notion of estimation by returning an **interval of values of θ compatible with the data**.
They are constructed based on pivotal quantities.

They have a dual relationship with hypothesis testing: an α-CR can be turned into a family of α-tests for $\theta \stackrel{?}{=} \theta_0$ and vice-versa.
In the rare cases in which we have most-powerful tests, we thus have associated Uniformly Most Accurate CI.

Multiple testing (NOT FOR EXAM)

Multiple Testing

Modern example: looking for signals in noise

- Interested in detecting presence of a signal $\mu(x_t)$, $t = 1, \ldots, T$ over a discretised domain, $\{x_1, \ldots, x_t\}$, on the basis of noisy measurements
- This is to be detected against some known background, say 0.
- May or may not be specifically interested in detecting the presence of the signal in some particular location x_t, but in detecting whether the signal is present anywhere in the domain.

Formally:

Does there exist a $t \in \{1, \ldots, T\}$ such that $\mu(x_t) \neq 0$?

or

for which t's is $\mu(x_t) \neq 0$?

Multiple Testing

More generally:

- Observe
$$Y_t = \mu(x_t) + \varepsilon_t, \qquad t = 1, \ldots, T.$$
- Wish to test, at some significance level α:
$$\begin{cases} H_0 : \mu(x_t) = 0 & \text{for all } t \in \{1, \ldots, T\}, \\ H_A : \mu(x_t) \neq 0 & \text{for some } t \in \{1, \ldots, T\}. \end{cases}$$
- May also be interested in which specific locations signal deviates from zero
- More generally: May have T hypotheses to test simultaneously at level α (they may be related or totally unrelated)
- Suppose we have a test statistic for each individual hypothesis $H_{0,t}$ yielding a p-value p_t.

Bonferroni Method

If we test each hypothesis individually, we will not maintain the level!

Can we maintain the level α?

Idea: use the same trick as for confidence regions!

Bonferroni

1. Test individual hypotheses separately at level $\alpha_t = \alpha / T$
2. Reject H_0 if at least one of the $\{H_{0,t}\}_{t=1}^{T}$ is rejected

Global level is bounded as follows:

$$\mathbb{P}[\cancel{H_0} | H_0] = \mathbb{P}\left[\bigcup_{t=1}^{T} \{\cancel{H_{0,t}}\} \middle| H_0 \right] \leq \sum_{t=1}^{T} \mathbb{P}[\cancel{H_{0,t}} | H_0] = T \frac{\alpha}{T} = \alpha$$

Holm-Bonferroni Method

- Advantage: Works for any (discrete domain) setup!
- Disadvantage: Too conservative when T large

Holm's modification increases average # of hypotheses rejected at level α (but does not increase power for overall rejection of $H_0 = \cap_{t \in T} H_{0,t}$)

Holm's Procedure

1. We reject $H_{0,t}$ for small values of a corresponding p-value, p_t
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. Starting from $t = 1$ and going up, reject all $H_{0,(t)}$ such that $p_{(t)}$ significant at level $\alpha/(T - t + 1)$. Stop rejecting at first insignificant $p_{(t)}$.

Genuine improvement over Bonferroni if want to detect as many signals as possible, not just existence of some signal

Both Holm and Bonferroni reject the global H_0 if and only if $\inf_t p_t$ significant at level α/T.

Taking Advantage of Structure: Independence

In the (special) case where individual test statistics are independent, one may use Sime's (in)equality,

$$\mathbb{P}\left[p_{(j)} \geq \frac{j\alpha}{T}, \text{ for all } j = 1, ..., T \middle| H_0 \right] \geq 1 - \alpha$$

(strict equality requires continuous test statistics, otherwise $\leq \alpha$)

Yields Sime's procedure (assuming independence)

1. Suppose we reject $H_{0,j}$ for small values of p_j
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. If, for some $j = 1, \ldots, T$ the p-value $p_{(j)}$ is significant at level $\frac{j\alpha}{T}$, then reject the global H_0.

Provides a test for the global hypothesis H_0, but does not "localise" the signal at a particular x_t

Taking Advantage of Structure: Independence

One can, however, devise a sequential procedure to "localise" Sime's procedure, at the expense of lower power for the global hypothesis H_0:

Hochberg's procedure (assuming independence)

1. Suppose we reject $H_{0,j}$ for small values of p_j
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. Starting from $j = T, T-1, \ldots$ and down, accept all $H_{0,(j)}$ such that $p_{(j)}$ insignificant at level $\alpha/(T-j+1)$.
4. Stop accepting for the first j such that $p_{(j)}$ is significant at level α/j, and reject all the remaining ordered hypotheses past that j going down.

Genuine improvement over Holm-Bonferroni both overall (H_0) and in terms of signal localisation:

1. Rejects "more" individual hypotheses than Holm-Bonferroni
2. Power for overall H_0 "weaker" than Sime's (for $T > 2$), much "stronger" than Holm (for $T > 1$).

Taking Advantage of Structure: Independence

Bonferroni, Hochberg, Simes