Statistics for Data Science: Week 6

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

Consider the simplest situation:

$$\Theta_0 = \{\theta_0\} \qquad \& \qquad \Theta_1 = \{\theta_1\}$$

The Neyman-Pearson Lemma - Continuous Case

Let $\boldsymbol{Y}$ have joint density/frequency $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad vs \qquad H_1 : f = f_1.$$

If $\Lambda(\boldsymbol{Y}) = f_1(\boldsymbol{Y})/f_0(\boldsymbol{Y})$ is a continuous random variable, then there exists a $k > 0$ such that

$$\mathbb{P}_0[\Lambda(\boldsymbol{Y}) \geq k] = \alpha$$

and the test whose test function is given by

$$\delta(\boldsymbol{Y}) = \mathbf{1}\{\Lambda(\boldsymbol{Y}) \geq k\},$$

is a *most powerful (MP)* test of H_0 versus H_1 at significance level α.

Proof.

Use obvious notation $\mathbb{E}_0$, $\mathbb{E}_1$, $\mathbb{P}_0$, $\mathbb{P}_1$ corresponding to H_0 or H_1. Let $G_0(t) = \mathbb{P}_0[\Lambda \le t]$. By assumption, G_0 is a differentiable distribution function, and so is onto $[0,1]$. Consequently, the set $\mathcal{K}_{1-\alpha} = \{t : G_0(t) = 1-\alpha\}$ is non-empty for any $\alpha \in (0,1)$. Setting $k = \inf\{t \in \mathcal{K}_{1-\alpha}\}$ we will have $\mathbb{P}_0[\Lambda \ge k] = \alpha$ and k is simply the $1-\alpha$ quantile of the distribution G_0. Consequently,

$$\mathbb{P}_0[\delta = 1] = \alpha \qquad (\text{since } \mathbb{P}_0[\delta = 1] = \mathbb{P}_0[\Lambda \ge k])$$

and therefore $\delta \in \mathcal{D}(\{\theta_0\}, \alpha)$ (i.e. δ indeed respects the level α).
To show that δ is also most powerful, it suffices to prove that if ψ is any function with $\psi(\boldsymbol{y}) \in \{0,1\}$, then

$$\mathbb{E}_0[\psi(\boldsymbol{Y})] \le \underbrace{\mathbb{E}_0[\delta(\boldsymbol{Y})]}_{=\alpha \text{(by first part of proof)}} \implies \underbrace{\mathbb{E}_1[\psi(\boldsymbol{Y})]}_{\beta_1(\psi)} \le \underbrace{\mathbb{E}_1[\delta(\boldsymbol{Y})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).

WLOG assume that f_0 and f_1 are density functions. Note that

$$f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}) \geq 0 \text{ if } \delta(\boldsymbol{y}) = 1 \quad \& \quad f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}) < 0 \text{ if } \delta(\boldsymbol{y}) = 0.$$

Therefore, since ψ can only take the values 0 or 1,

$$\begin{aligned} \psi(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) &\leq \delta(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) \\ \int_{\mathbb{R}^n} \psi(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) d\boldsymbol{y} &\leq \int_{\mathbb{R}^n} \delta(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) d\boldsymbol{y} \end{aligned}$$

Rearranging the terms yields

$$\begin{aligned} \int_{\mathbb{R}^n} (\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})) f_1(\boldsymbol{y}) d\boldsymbol{y} &\leq k \int_{\mathbb{R}^n} (\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})) f_0(\boldsymbol{y}) d\boldsymbol{y} \\ \implies \mathbb{E}_1[\psi(\boldsymbol{Y})] - \mathbb{E}_1[\delta(\boldsymbol{Y})] &\leq k \left(\mathbb{E}_0[\psi(\boldsymbol{y})] - \mathbb{E}_0[\delta(\boldsymbol{Y})] \right) \end{aligned}$$

But $k > 0$ by assumption, so when $\mathbb{E}_0[\psi(\boldsymbol{Y})] \leq \mathbb{E}_0[\delta(\boldsymbol{Y})]$ the RHS is negative, i.e. δ is an MP test of H_0 vs H_1 at level α. □

- Basically we reject if the likelihood of θ_0 is k times higher than the likelihood of θ_1. This is called a likelihood ratio test, and Λ is the likelihood ratio statistic: *how much more plausible is the alternative than the null?*

- When Λ is a continuous RV, the choice of k is essentially unique. That is, if k' is such that $\delta' = \mathbf{1}\{\Lambda \geq k'\} \in \mathcal{D}(\{\theta_0\}, \alpha)$, then $\delta = \delta'$ almost surely.

- The resulting most powerful test is not necessarily unique.

- Unless Λ is continuous, the most powerful test is not necessarily guaranteed to exist.

- The problem if Λ is a RV with a discontinuous dist is that there may exist no k for which the equation $\mathbb{P}_0[\Lambda \geq k] = \alpha$ has a solution.

- In any case, typically the distribution of the test statistic converges to a continuous limit with large n, so these problems become inessential.

Example (Poisson Distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Poisson(\mu)$ and for $\mu_1 > \mu_0$ consider the hypotheses:

$$H_0 : \mu = \mu_0 \qquad vs \qquad H_1 : \mu = \mu_1.$$

These correspond to the Higgs example by setting $\mu_0 = b$ and $\mu_1 = b + s$.

Applying the Neyman-Pearson lemma gives a test statistic

$$\delta(Y_1, \ldots, Y_n) = \mathbf{1}\left\{\sum_{i=1}^{n} Y_i > q_{1-\alpha}\right\},$$

provided α is such that $G_0(q_{1-\alpha}) = \mathbb{P}_{\mu_0}[\tau(Y_1, \ldots, Y_n) \leq q_{1-\alpha}] \overset{!}{=} 1 - \alpha$.

Since the Y_i are independent, one can easily show that

$$\tau(Y_1, \ldots, Y_n) \overset{H_0}{\sim} Poisson(n\mu_0).$$

This being a discrete distribution, the only α for which we get an MP test are

$$e^{-n\mu_0}, e^{-n\mu_0}\left(1 + n\mu_0\right), e^{-n\mu_0}\left(1 + n\mu_0 + \frac{(n\mu_0)^2}{2}\right), \ldots \text{and so on}$$

Nevertheless notice that as $n \to \infty$, these values become dense near the origin.

When $\{\Theta_0, \Theta_1\}$ are not singletons, choosing a most powerful test is a much stronger requirement:

1. It should respect the level for all $\theta \in \Theta_0$, i.e.

$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \{\delta : \mathcal{Y}^n \to \{0,1\} : \mathbb{E}_\theta[\delta] \leq \alpha, \ \forall\, \theta \in \Theta_0\}$$

2. It should be most powerful for all $\theta \in \Theta_1$ (i.e. for all possible simple alternatives),

$$\mathbb{E}_\theta[\delta] \geq \mathbb{E}_\theta[\delta'] \qquad \forall \theta \in \Theta_1 \quad \& \quad \delta' \in \mathcal{D}(\Theta_0, \alpha)$$

Unfortunately UMP tests rarely exist. Why?
$\hookrightarrow$ Consider $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- A UMP test must be MP test for any $\theta \neq \theta_1$.
- But the form of the MP test typically differs for $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$!
  $\hookrightarrow$ e.g. recall exponential mean example

Example (No UMP test exists)

Let $Y_1, ..., Y_n \sim$Bernoulli(θ) and suppose we want to test:

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

at some level α. To this aim, consider first

$$H_0' : \theta = \theta_0 \qquad vs \qquad H_1' : \theta = \theta_1$$

Neyman-Pearson lemma gives test statistics

$$T = \frac{f(\boldsymbol{Y}; \theta_1)}{f(\boldsymbol{Y}; \theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_{i=1}^n Y_i}$$

- If $\theta_1 > \theta_0$ then T increasing in $\sum_{i=1}^n Y_i$
  - $\hookrightarrow$ MP test would reject for large values of $\sum_{i=1}^n Y_i$
- If $\theta_1 < \theta_0$ then T decreasing in $\sum_{i=1}^n Y_i$
  - $\hookrightarrow$ MP test would reject for small values of $\sum_{i=1}^n Y_i$

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

- One sided hypotheses: when Θ_0 is a an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.
    - For example, in one-parameter exponential families, one simply uses the Neyman-Pearson lemma, taking the null to be $\theta = \theta_0$ and the alternative $\theta = \theta_1$ for any $\theta_1 \in \Theta_1$ (the form of the test depends only on the direction of the null and the boundary of the null).
    - This generalises to families admitting a so-called "monotone likelihood ratio"
    - In the absence of the "monotone likelihood ratio" property, one can seek locally most powerfull tests, near the hypothesis boundary. It can be shown that the score function (derivative of the logliklihood) at the boundary θ_0 can serve as a test statistic to this aim.
- General hypothesis pairs: we need to abandon optimality, and search for sensible tests. But the likelihood ratio idea can serve us well in this pursuit.

Consider now the multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Maximum Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio statistic* corresponding to the pair of hypotheses $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$ is defined to be

$$\Lambda(\boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{Y}; \theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} f(\boldsymbol{Y}; \theta)} = \frac{\sup_{\boldsymbol{\theta} \in \Theta} L(\theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\theta)}$$

- Intuition: choose the "most favourable" $\theta \in \Theta_0$ (in favour of H_0) and compare it against the "most favourable" $\theta \in \Theta_1$ (in favour of H_1) in a simple vs simple setting (applying NP-lemma)
- Typically Θ_0 is a lower dimensional subspace of Θ_1, so taking sup over Θ (rather than Θ_1) incurs no loss. In this case $\Theta_0 \cap \Theta_1 \neq \emptyset$, but $Leb(\Theta_0 \cap \Theta_1) = 0$, which suffices.

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0$$

$$\Lambda(\boldsymbol{Y}) = \frac{\sup_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+} f(\boldsymbol{Y};\mu,\sigma^2)}{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}^+} f(\boldsymbol{Y};\mu,\sigma^2)} = \left(\frac{\sum_{i=1}^n (Y_i - \mu_0)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}\right)^{\frac{n}{2}} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}}$$

So reject when $\Lambda \geq k$, where k is s.t. $\mathbb{P}_0[\Lambda \geq k] = \alpha$. Distribution of Λ? By monotonicity look only at

$$\begin{aligned} \frac{\sum_{i=1}^n (Y_i - \mu_0)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} &= 1 + \frac{n(\bar{Y} - \mu_0)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 + \frac{1}{n-1}\left(\frac{n(\bar{Y} - \mu_0)^2}{S^2}\right) \\ &= 1 + \frac{T^2}{n-1} \end{aligned}$$

With $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ and $T = \sqrt{n}(\bar{Y} - \mu_0)/S \overset{H_0}{\sim} t_{n-1}$.

So $T^2 \overset{H_0}{\sim} F_{1,n-1}$ and k may be chosen appropriately.

Example

Let $Y_1, ..., Y_m \overset{iid}{\sim} \text{Exp}(\lambda)$ and $Z_1, ..., Z_n \overset{iid}{\sim} \text{Exp}(\theta)$. Assume $\boldsymbol{Y}$ indep $\boldsymbol{Z}$.

$$\text{Consider:} \quad H_0 : \theta = \lambda \qquad \text{vs} \qquad H_1 : \theta \neq \lambda$$

i.e. $(\theta, \lambda) \in R_+^2$ against $(\theta, \lambda) \in$ 45 degree line

$$\underset{\sup_{(\lambda,\theta)\in\mathbb{R}_+^2} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)}{\text{Unrestricted MLEs:}} \quad \hat{\lambda} = 1/\bar{Y} \qquad \& \qquad \hat{\theta} = 1/\bar{Z}$$

$$\underset{\sup_{(\lambda,\theta)\in\{(y,z)\in\mathbb{R}_+^2 : y=z\}} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)}{\text{Restricted MLEs:}} \quad \hat{\lambda}_0 = \hat{\theta}_0 = \left[\frac{m\bar{Y} + n\bar{Z}}{m+n}\right]^{-1}$$

$$\implies \Lambda = \left(\frac{m}{m+n} + \frac{n}{n+m}\frac{\bar{Z}}{\bar{Y}}\right)^m \left(\frac{n}{n+m} + \frac{m}{m+n}\frac{\bar{Y}}{\bar{Z}}\right)^n$$

Depends on $T = \bar{Y}/\bar{Z}$ and can make Λ large/small by varying T.

$\hookrightarrow$ But $T \overset{H_0}{\sim} F_{2m,2n}$ so given α we may find the critical value k.

More often than not, dist(Λ) intractable
$\hookrightarrow$(and no simple dependence on T with tractable distribution either)

Consider asymptotic approximations?

Setup

- Θ open subset of $\mathbb{R}^p$
- either $\Theta_0 = \{\boldsymbol{\theta}_0\}$ or Θ_0 open subset of $\mathbb{R}^s$, where $s < p$
- Concentrate on $\boldsymbol{Y} = (Y_1, ..., Y_n)$ has iid components.
- Initially restrict attention to $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. LR becomes:

$$\Lambda_n(\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{f(Y_i; \hat{\boldsymbol{\theta}}_n)}{f(Y_i; \boldsymbol{\theta}_0)}$$

  where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$.
- Impose regularity conditions from MLE asymptotics

Theorem (Wilks' Theorem, case $p = 1$)

Let $Y_1, \ldots, Y_n$ be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}$ and satisfying conditions (A1)-(A6), with $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$. If the MLE sequence $\hat{\theta}_n$ is consistent for θ, then the likelihood ratio statistic Λ_n for $H_0 : \theta = \theta_0$ satisfies

$$2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_1$$

when H_0 is true.

- Obviously, knowing approximate distribution of $2 \log \Lambda_n$ is as good as knowing approximate distribution of Λ_n for the purposes of testing (by monotonicity and rejection method).
- Theorem extends immediately and trivially to the case of general p and for a hypothesis pair $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.
  (i.e. when null hypothesis is simple)

Proof ($*$).

Under the conditions of the theorem and when H_0 is true,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta_0))$$

Now take logarithms and expand in a Taylor series around $\hat{\theta}_n$,

$$\begin{aligned}
\log \Lambda_n &= \sum_{i=1}^{n} [\ell(Y_i; \hat{\theta}_n) - \ell(Y_i; \theta_0)] = \sum_{i=1}^{n} [\ell(Y_i; \hat{\theta}_n) - \ell(Y_i; \hat{\theta}_n)] + \\
&\quad + (\theta_0 - \hat{\theta}_n) \sum_{i=1}^{n} \ell'(Y_i; \hat{\theta}_n) - \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \sum_{i=1}^{n} \ell''(Y_i; \theta_n^*) \\
&= -\frac{1}{2} n(\hat{\theta}_n - \theta_0)^2 \frac{1}{n} \sum_{i=1}^{n} \ell''(Y_i; \theta_n^*)
\end{aligned}$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0.

If H_0 is true, and since $\hat{\theta}_n$ is a consistent sequence, θ_n^* is sandwiched so

$$\theta_n^* \xrightarrow{p} \theta_0.$$

Hence under assumptions (A1)-(A6), and when H_0 is true, a first order Taylor expansion about θ_0, the continuous mapping theorem and the LLN give

$$\frac{1}{n}\sum_{i=1}^{n} \ell''(Y_i; \theta_n^*) \xrightarrow{p} -\mathbb{E}_{\theta_0}[\ell''(Y_i; \theta_0)] = \mathcal{I}_1(\theta_0)$$

On the other hand, by the continuous mapping theorem,

$$n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{V}{\mathcal{I}_1(\theta_0)}$$

Applying Slutsky's theorem now yields the result. □ □

Theorem (Wilk's theorem, general p, general $s \leq p$)

Let $Y_1, ..., Y_n$ be iid random variables with density (frequency) depending on $\boldsymbol{\theta} \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $\mathcal{I}_1(\boldsymbol{\theta}) = \mathcal{J}_1(\boldsymbol{\theta})$. If the MLE sequence $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$, then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_s$ when H_0 is true.

Comments:

- Note that it may potentially be that $s < p$, and this is accommodated by the theorem
- Hypotheses of the form $H_0 : \{g_j(\boldsymbol{\theta}) = a_j\}_{j=1}^s$, for g_j differentiable real functions, can also be handled by Wilks' theorem:
  - Define $(\phi_1, ..., \phi_p) = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), ..., g_p(\boldsymbol{\theta}))$
  - $g_{s+1}, ..., g_p$ defined so that $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is 1-1
  - Apply theorem with parameter ϕ

Many other tests possible. For example:

- Wald's test
  - ↪ For a simple null, may compare the unrestricted MLE with the MLE under the null. Large deviations indicate evidence against null hypothesis. Distributions are approximated for large n via the asymptotic normality of MLEs.
- Score Test
  - ↪ For a simple null, if the null hypothesis is false, then the loglikelihood gradient at the null should not be close to zero, at least when n reasonably large: so measure its deviations form zero. Use asymptotics for distributions (under conditions we end up with a χ^2)
- ...

So far focussed on 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 $\delta^*(\boldsymbol{y}) = 1$.

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 $\boldsymbol{y}$ some of these rules reject the null, while others do not
2. Which is the smallest α for which H_0 is rejected given $\boldsymbol{y}$?

Definition (p–Value)

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

$$\alpha_1 < \alpha_2 \implies \{\boldsymbol{y} \in \mathcal{Y}^n : \delta_{\alpha_1}(\boldsymbol{y}) = 1\} \subseteq \{\boldsymbol{y} \in \mathcal{Y}^n : \delta_{\alpha_2}(\boldsymbol{y}) = 1\}.$$

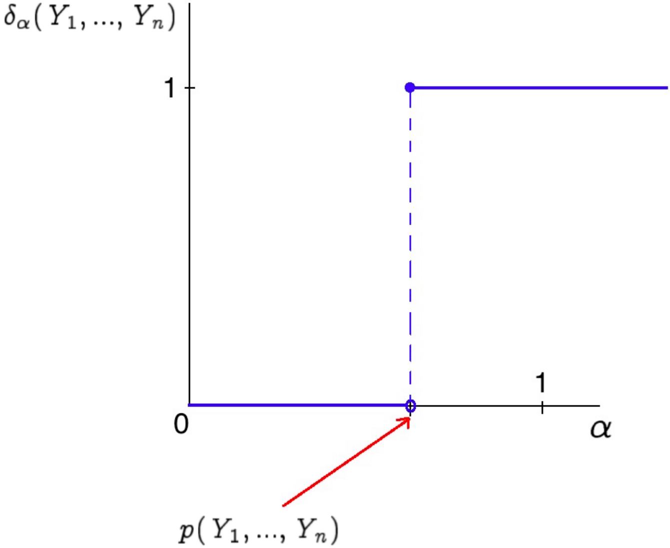
The p–value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\boldsymbol{y}) = \inf\{\alpha : \delta_\alpha(\boldsymbol{y}) = 1\}$$

$\hookrightarrow$ The p–value is the smallest value of α for which the null would be rejected at level α, given $\boldsymbol{Y} = \boldsymbol{y}$.

Most usual setup:

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



$\delta_\alpha(Y_1, ..., Y_n)$
1
1
0
α
$p(Y_1, ..., Y_n)$

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
  - ↪ Small p–value provides evidence against H_0
  - ↪ Large p–value provides no evidence against H_0
- How small does "small" mean?
  - ↪ 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, then we have witnessed something quite unusual under the null hypothesis
- Gives evidence against the null hypothesis

Example (Normal Mean)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ 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, $T = \sqrt{n}\bar{Y}/S \overset{H_0}{\sim} t_{n-1}$.

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

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

Consider two samples (datasets),

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

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

Obtain $p(\boldsymbol{y}) = 0.32$ while $p(\boldsymbol{y}') = 0.006$.

- Reporting a p–value does not necessarily mean making a decision
- A small p–value can simply reflect our "confidence" in rejecting a null

Recall example: Statisticians working for Trump gather iid sample $\boldsymbol{Y}$ from Florida with $Y_i = \mathbf{1}\{\text{vote Biden}\}$. Trumps team want to test

$$\begin{cases} H_0 : & \text{Trump wins Florida} \\ H_1 : & \text{Biden wins Florida} \end{cases}$$

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

What if statisticians working for newspaper, not Trump?

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

A Glance Back at Point Estimation

- Let $Y_1, ..., Y_n$ 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 $\hat{P}[\text{Biden wins}] = 0.52$ can actually see what p–value we get when testing $H_0 : P[\text{Biden wins}] \geq 1/2$.
- 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?

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)

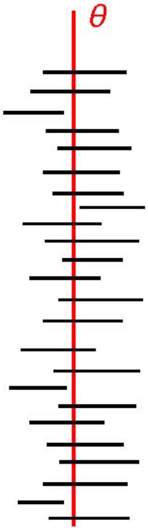
Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be random variables with joint distribution depending on $\theta \in \mathbb{R}$ and let $L(\boldsymbol{Y})$ and $U(\boldsymbol{Y})$ be two statistics with $L(\boldsymbol{Y}) < U(\boldsymbol{Y})$ a.s. Then, the random interval $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ is called a $100(1-\alpha)\%$ confidence interval for θ if

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

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!

- 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 $\boldsymbol{Y} = \boldsymbol{y}$, the interval $[L(\boldsymbol{y}), U(\boldsymbol{y})]$ 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 (The example that says all)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Then $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$, so that

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

and since

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

we obviously have

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

So that the random interval $[L(\boldsymbol{Y}), U(\boldsymbol{Y})] = \left[\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}} \right]$ is a 95% confidence interval for μ.
Central Limit Theorem: same argument can yield approximate 95% CI when $Y_1, ..., Y_n$ are iid, $\mathbb{E} Y_i = \mu$ and $\mathrm{var}(Y_i) = 1$, regardless of their distribution.

Example (continued)

Notice that the interval is centred at $\bar{Y}$, the MLE of μ. It's often thus written:

$$\boxed{\bar{Y} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

Observations:

- The length of the interval is $2z_{1-\alpha/2}\sigma/\sqrt{n}$, which depends on σ^2, n and α.
- The parameter σ^2 is beyond our control.
- We can nevertheless control n and $1-\alpha$. Increasing n, the length of the interval decays as $1/\sqrt{n}$.
- Reducing α (i.e. increasing $1-\alpha$) increases the length of the interval (the dependence is quite non-linear, and 5% is chosen as a "sweet spot").

What can we learn from previous example?

Definition (Pivot)

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

$\hookrightarrow \sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$ 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

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

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

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. Recall that MLE $\hat{\theta}$ is $\hat{\theta} = Y_{(n)}$, with distribution

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

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

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

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

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

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

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. $\to$ Pivotal quantity should now be function $g(\boldsymbol{Y}; \theta_k)$ which

1. Depends on $\boldsymbol{Y}$, 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

$\to$ 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 $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$.

What about higher dimensional parameters?

Definition (Confidence Region)

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be random variables with joint distribution depending on $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. A random subset $R(\boldsymbol{Y})$ of Θ depending on $\boldsymbol{Y}$ is called a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ if

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

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

- No restriction requiring $R(\boldsymbol{Y})$ 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

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

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

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

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

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

Example

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

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

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

Example (cont'd)

Consider the random variable

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

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

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

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

Visualisation of high-dimensional CR's can be hard

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

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

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

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

Bonferroni's inequality implies that

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

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

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

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

Suppose $R(\boldsymbol{Y})$ 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:

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

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

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

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 $\delta(\boldsymbol{Y}; \boldsymbol{\theta}_0)$ is appropriate test function for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

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

Coverage probability of $R^*(\boldsymbol{Y})$ is

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

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

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

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 a 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 $\mu(x_t) \neq 0$?

More generally:

- Observe

$$Y_t = \mu(x_t) + \varepsilon_t, \qquad t = 1, \dots, T.$$

- Wish to test, at some significance level α:

$$\begin{cases} H_0 : \mu(x_t) = 0 & \text{for all } t \in \{1, \dots, T\}, \\ H_A : \mu(x_t) \neq 0 & \text{for some } t \in \{1, \dots, T\}. \end{cases}$$

- 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:

$$\mathbb{P}[\text{not } H_0 | H_0] = \mathbb{P}\left[\bigcup_{t=1}^{T}\{\text{not } H_{0,t}\}\middle| H_0\right] \leq \sum_{t=1}^{T}\mathbb{P}[\text{not } H_{0,t} | H_0] = T\frac{\alpha}{T} = \alpha$$

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

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 $T > 1$).

Bonferroni, Hochberg, Simes