# Hypothesis Testing

Sofia Olhede

EPFL

October 19, 2020

1. Hypothesis Testing

# Hypothesis Testing IX

reminder:
$H_0$ : NULL hypothesis
$H_1$: alternative hypothesis

- Neyman-Pearson setup naturally exploits any asymmetric structure.
- But, if natural asymmetry absent, need judicious choice of $H_0$.
- Consider simplest situation: $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.
  $H_0$ $H_1$
- The Neyman Pearson Lemma: Let $\boldsymbol{Y}$ have joint density/frequency $f$ where $f \in \{f_0, f_1\}$. We wish to test

"$f_1$ evaluated at the obs. $Y$"

$$H_0 : \quad f = f_0 \quad \text{and} \quad f = f_1.$$

- If $\Lambda(\mathrm{Y}) = f_1(\mathrm{Y})/f_0(\mathrm{Y})$ is a continuous random variable, then there exists a $k > 0$ such that $\Pr\{\Lambda(\mathrm{Y}) \geq k | H_0\} = \alpha$ and the test whose test function is given by $\delta(\mathrm{Y}) = \mathrm{I}(\Lambda(\mathrm{Y}) \geq k)$ is a most powerful (MP) test of $H_0$ versus $H_1$ at significance level $\alpha$.

putting it simply:
if $\Lambda(Y)$ is continuous, we already found the test.
No need to search more

test function

indicator func : $\begin{cases} 0 & \text{if } \Lambda(Y) < k \\ 1 & \text{if } \Lambda(Y) \geq K \end{cases}$

Proof:

- We will index both the expectation and the probability measure by the distribution under either the null or the alternative hypothesis. ?
- We denote by $G_0(t) = \text{Pr}_0\{\Lambda \leq t\}$.
- We have assumed $G_0$ is a differentiable distribution function. It is therefore onto $[0, 1]$.
- We can deduce that 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 $\text{Pr}_0\{\Lambda \geq k\} = \alpha$ and $k$ is simply the $1 - \alpha$ quantile of the distribution $G_0$.
- Thus it follows

$$\text{Pr}_0\{\delta = 1\} = \alpha, \quad (\text{as} \quad \text{Pr}_0\{\delta = 1\} = \text{Pr}\{\Lambda \geq k\}.)$$

- Therefore it follows that $\delta$ respects the levels of $\alpha$.

- To show that $\delta$ is also most powerful, it suffices to prove that if $\psi$ is ? any function with $\psi(\boldsymbol{y}) \in \{0,1\}$ then

$$\mathbb{E}_0\{\psi(\boldsymbol{Y})\} \leq \mathbb{E}_0\{\delta(\boldsymbol{Y})\} = \alpha \Rightarrow \quad \mathbb{E}_1\{\psi(\boldsymbol{Y})\} \leq \mathbb{E}_1\{\delta(\boldsymbol{Y})\}. \qquad (1)$$

This orders $\beta_1(\psi)$ relative to $\beta_1(\delta)$.

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

$$f_1(\boldsymbol{y}) - kf_0(\boldsymbol{y}) \geq 0 \quad \text{if } \delta(\boldsymbol{y}) = 1, \quad \& \quad f_1(\boldsymbol{y}) - kf_0(\boldsymbol{y}) < 0 \quad \text{if } \delta(\boldsymbol{y}) = 0. \qquad (2)$$

Since $\psi$ can only take the values 0 or 1.

$$\psi(\boldsymbol{y})\{f_1(\boldsymbol{y}) - kf_0(\boldsymbol{y})\} \leq \delta(\boldsymbol{y})\{f_1(\boldsymbol{y}) - kf_0(\boldsymbol{y})\}$$
$$\int_{\mathbb{R}^n} \psi(\boldsymbol{y})\{f_1(\boldsymbol{y}) - kf_0(\boldsymbol{y})\}\, d\boldsymbol{y} \leq \int_{\mathbb{R}^n} \delta(\boldsymbol{y})\{f_1(\boldsymbol{y}) - kf_0(\boldsymbol{y})\}\, d\boldsymbol{y}.$$

Rearranging the terms yields:

?

- that is

$$\int_{\mathbb{R}^n} \{\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})\} f_1(\boldsymbol{y})\, d\boldsymbol{y} \leq k \int_{\mathbb{R}^n} \{\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})\} f_0(\boldsymbol{y})\, d\boldsymbol{y}$$
$$\mathbb{E}_1\{\psi(\boldsymbol{Y})\} - \mathbb{E}_1\{\delta(\boldsymbol{Y})\} \leq k\{\mathbb{E}_0\{\psi(\boldsymbol{Y})\} - \mathbb{E}_0\{\delta(\boldsymbol{Y})\}\}.$$

- However as $k > 0$ (by assumption). Thus when $\mathbb{E}_0\{\psi(\boldsymbol{Y})\} < \mathbb{E}_0\{\delta(\boldsymbol{Y})\}$ then the RKS is negative then $\delta$ is an MP test of $H_0$ vs $H_1$ at level $\alpha$.

- Basically we reject if the likelihood of $\theta_0$ is $k$ times higher than the likelihood of $\theta_1$. This is called a likelihood ratio test, and $\Lambda$ is the likelihood ratio statistic: how much more plausible is the alternative than the null?

random var

- When $\Lambda$ is a continuous RV, the choice of $k$ is essentially unique. That is, if $k'$ is such that $\delta' = \mathrm{I}(\Lambda \geq k') \in \mathcal{D}(\{\theta_0\}, \alpha)$ then $\delta = \delta'$ almost surely.
- The resulting most powerful test is not necessarily unique.
- Unless $\Lambda$ is continuous, the most powerful test is not necessarily guaranteed to exist.

Here's why

- The problem if $\Lambda$ is a RV with a discontinuous dist is that there may exist no $k$ for which the equation $\Pr_0\{\Lambda \geq k\} = \alpha$ 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: the Poisson distribution.
- Let $Y_1 \dots Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\mu)$ and for $\mu_1 > \mu_0$ consider the hypotheses:

$$H_0: \ \mu = \mu_0 \quad \text{vs} \quad H_1: \ \mu = \mu_1.$$

- Applying the Neyman-Pearson lemma gives a test statistic

$T = \frac{f(Y;\theta_1)}{f(Y;\theta_0)} = \frac{\theta_1^n e^{-\theta_1/n}}{\theta_2^n e^{-\theta_2/n}} = ?$

$$\delta(Y_1, \dots, Y_n) = \mathrm{I}\{\underbrace{\sum_i Y_i}_{\text{sum of the Poisson}} > q_{1-\alpha}\},$$

if $\alpha$ is such that $G_0(q_{1-\alpha}) = \Pr\{\tau(Y_1, \dots, Y_n) \le q_{1-\alpha}\} = 1 - \alpha$.

- Since the $Y_i$ are independent random variables, one can show that

"using moment generating functions, we can show that under the null hypothesis, we have a poisson but with a quicken rate."

$$\tau(Y_1, \dots, Y_n) \overset{H_0}{\sim} \text{Poisson}(\underbrace{n\mu_0}_{\text{null hypothesis}}).$$

quicken rate

?

- This being a discrete distribution, the only $\alpha$ for which we get an MP test are $e^{-n\mu_0}$, $e^{-n\mu_0}\{1 + n\mu_0\}$, $e^{-n\mu_0}\{1 + n\mu_0 + \frac{1}{2}(n\mu_0)^2\}$, and so on. Nevertheless notice that as $n \to \infty$ these values become dense near the origin.

is a single value $\mathbb{R}$ ?

- When $\{\Theta_0, \Theta_1\}$ are are not singletons, choosing a most powerful test is a much stronger requirement: ?
- It should respect the level for all $\theta \in \Theta_0$ that is

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

- It should be most powerful for all $\theta \in \Theta_1$ (e.g. the set of alternatives):

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

uniformly most powerful

- Unfortunately UMP tests rarely exist. Why?
- Consider the simple test of $H_0 \ \theta = \theta_0$ versus $H_1 \ \theta \neq \theta_0$.
- 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 $\theta_1 < \theta_0$. For example consider the exponential mean example.

- Example of non-existence of UMP.
- Let $Y_1, \ldots, Y_n \sim \mathrm{Bernoulli}(\theta)$. We want to test:

?

$\gamma$

$$H_0: \quad \theta = \theta_0, \quad \text{vs} \quad H_1: \ \theta \neq \theta_0,$$

at some given level $\alpha$. Let us start with the simpler problem of testing

$$H_0: \quad \theta = \theta_0, \quad \text{vs} \quad H_1: \ \theta = \theta_1.$$

- We start from the Neyman–Pearson lemma which gives as the test statistic

$= \frac{\theta_1^n (1-\theta_1)^{1-n}}{\theta_2^n (1-\theta_2)^{1-n}} = ?$

doesn't depend on the data (Y)

depends on the data

$$T = \frac{f(\mathrm{Y};\theta_1)}{f(\mathrm{Y};\theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_i Y_i}.$$

- If $\theta_1 > \theta_0$ then $T$ is increasing in $\sum_i Y_i$. Thus MP test would reject for large values of $\sum_i Y_i$.
- If $\theta_1 < \theta_0$ then $T$ is decreasing in $\sum_i Y_i$. Thus MP test would reject for small values of $\sum_i Y_i$.

- So what can we do for more general $\{\Theta_0, \Theta_1\}$?
- One sided tests. When $\Theta_0$ is in the form of an interval $(-\infty, \theta_0]$ or $[\theta_0, \infty)$ and $\Theta_1$ is its complement, then 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 an alternative of $\theta = \theta_1$ 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 $\theta_0$ 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.

The supremum (sup; plural suprema) of a subset S of a partially ordered set T is the least element in T that is greater than or equal to all elements of S, if such an element exists. Consequently, the supremum is also referred to as the least upper bound (or LUB).

- Consider now the multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$ with general $\Theta_0$ and $\Theta_1$.
  i.e. not singletons anymore
- As noted optimality breaks down. → "we can't get simple results as the NP by comparing one density with another"
- But we can still seek general-purpose approaches.
- Definition: Likelihood ratio. which is
  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

$$\Lambda(\boldsymbol{\theta}) = \frac{\sup_{\theta \in \Theta_1} f(\boldsymbol{Y}; \boldsymbol{\theta})}{\underbrace{\sup_{\theta \in \Theta_0}}_{\text{best } \theta \text{ (single value) from } \Theta_0} f(\boldsymbol{Y}; \boldsymbol{\theta})} = \frac{\sup_{\theta \in \Theta_1} L(\boldsymbol{\theta})}{\sup_{\theta \in \Theta_0} L(\boldsymbol{\theta})}.$$

ie same as applying NP to a single value

- Intuition: choose the "most favourable" $\boldsymbol{\theta} \in \Theta_0$ (in favour of $H_0$) and compare to the "most favourable" $\boldsymbol{\theta} \in \Theta_1$ (in favour of $H_1$), in a simple vs simple setting (applying NP-lemma) → i.e. the params of max likelihood = log f
- Typically $\Theta_0$ is a a lower dimensional subspace of $\Theta_1$ so taking sup over $\Theta$ rather than $\Theta_1$ incurs no loss.

- Example: Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. We want to test:

because $Y \overset{iid}{\sim} N(\mu, \sigma^2)$ we can factorize the likelihood

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

at some given level $\alpha$. Assume both parameters unknown.

- Then $f(Y_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{Y_i-\mu}{\sigma}\right)^2\right)$ so $\ell(Y_i; \mu, \sigma^2) = \sum_i^n \left(\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y_i-\mu)^2\right)$

$$\Lambda = \frac{\sup_{\mu,\sigma^2} f(\boldsymbol{Y}; \mu, \sigma^2)}{\sup_{\mu=\mu_0} f(\boldsymbol{Y}; \mu, \sigma^2)} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{n/2} = \left(\frac{\sum_i (Y_i - \mu_0)^2}{\sum_i (Y_i - \bar{Y})^2}\right)^{n/2}.$$

see video

- So reject when $\Lambda > k$ where $k$ is found from the null distribution. By the monotonicity we only need to compare

T-square distribution

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

- With $S^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$ 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 $\boldsymbol{Z}$.

Consider: $H_0 : \theta = \lambda$ vs $H_1 : \theta \neq \lambda$

Unrestricted MLEs: $\hat{\lambda} = 1/\bar{Y}$ & $\hat{\theta} = 1/\bar{Z}$

$\sup_{(\lambda,\theta)\in\mathbb{R}^2_+} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)$

weighted mean of Y and Z of lengths m and n

means $Y_i$ and $Z_i$ have the same distributions

Restricted MLEs: $\hat{\lambda}_0 = \hat{\theta}_0 = \left[\frac{m\bar{Y} + n\bar{Z}}{m+n}\right]^{-1}$

$\sup_{(\lambda,\theta)\in\{(y,z)\in\mathbb{R}^2_+ : y=z\}} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)$

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

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

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

F distribution

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

Consider asymptotic approximations?

Setup

Likelihood (?) Ratio test

- $\Theta$ open subset of $\mathbb{R}^p$
- either $\Theta_0 = \{\boldsymbol{\theta}_0\}$ or $\Theta_0$ 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:

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

→ estimator under MLE
→ estimator under constant

?

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

the MLE conditions

for a simple Null hypothesis

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

*Let $Y_1, \dots, Y_n$ 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 $\theta$, then the likelihood ratio statistic $\Lambda_n$ 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 $\Lambda_n$ 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 $\theta_n^*$ lies between $\hat{\theta}_n$ and $\theta_0$.

If $H_0$ is true, and since $\hat{\theta}_n$ is a consistent sequence, $\theta_n^*$ 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 $\theta_0$, 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. $\square$