Hypothesis Testing Cont'd

Sofia Olhede

EPFL

October 26, 2020

Likelihood ratio test

- Theorem (Wilks theorem for general $s < p$): Let $Y_1, \ldots, Y_n$ be iid random variables with density (frequency) depending on $\boldsymbol{\theta} \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $\mathcal{I}_1(\boldsymbol{\theta}) = \mathscr{I}_1(\boldsymbol{\theta})$. If the MLE sequence $\widehat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi_s^2$ when H_0 is true.
- Note that it may potentially be that $s < p$, and this is accommodated by the theory,
- 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 $\left(\phi_1, \ldots \phi_p\right) = g(\boldsymbol{\theta}) = \left(g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta})\right)$.
- $g_{s+1}, \ldots, g_p$ defined so that $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is 1-1.
- Apply theorem with parameter ϕ.

Likelihood ratio test

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).

The infamous p-value

- Fix a significance level α for the test;
- Consider rules δ respecting this significance level
  We choose one of those rules, δ^*, based on power considerations;
- 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)
- 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.
- Fix a family $\{\delta_\alpha\}_{\alpha \in (0,1)}$ of decision rules, with δ_α having level α.
- For a given $\boldsymbol{y}$ some of these rules reject the null, while others do not.
- Which is the smallest α for which H_0 is rejected given $\boldsymbol{y}$?

The infamous p-value

- Let $\{\delta_\alpha\}_\alpha$ be a family of test functions satisfying

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

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

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

- The p–value is the smallest value of α for which the null would be rejected at level α, given $\boldsymbol{Y} = \boldsymbol{y}$.
- The most usual setup:
  * Have a single test statistic T
  * Construct family $\delta_\alpha(\boldsymbol{y}) = \mathrm{I}\{T(\boldsymbol{y}) > k_\alpha\}$.
  * If $\Pr_{H_0}\{T \leq t\} = G(t)$ then
  $p(\boldsymbol{y}) = \Pr_{H_0}\{T(\boldsymbol{Y}) \geq T(\boldsymbol{y})\} = 1 - G(T(\boldsymbol{y}))$.

The infamous *p*-value

- Notice: contrary to Neyman Pearson-framework did not make explicit decision!
- We simply report 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 problem).
- 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.

Normal mean

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

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

- Likelihood ratio test: reject when T^2 large $T = \sqrt{n}\overline{Y}/S \overset{H_0}{\sim} t_{n-1}$.
- Since $T^2 \overset{H_0}{\sim} F_{1,n-1}$ p–value is

$$p(\boldsymbol{y}) = \Pr_{H_0}\{T^2(\boldsymbol{Y} \geq T^2(\boldsymbol{y})\} = 1 - G_{F_{1,n-2}}(T^2(\boldsymbol{y})).$$

- Consider two samples (data sets)

$$\boldsymbol{y} = \begin{pmatrix} 0.66 & 0.28 & -0.99 & 0.007 & -0.29 & -1.88 & -1.24 & 0.94 & 0.53 & -1.2 \end{pmatrix}.$$

$$\boldsymbol{y} = \begin{pmatrix} 1.4 & 0.48 & 2.86 & 1.02 & -1.38 & 1.42 & 2.11 & 2.77 & 1.02 & 1.87 \end{pmatrix}.$$

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

Normal mean

- Reporting a p–value does not necessarily mean making a decision.
- A small p–value can simply reflect our “confidence” in rejecting a null.
- A Glance Back at Point Estimation.
- Let $Y_1, \ldots, Y_n$ be iid random variables with density (frequency) $f(\cdot;\theta)$.
- Problem with point estimation: $\Pr_\theta\{\widehat{\theta} = \theta\}$ typically small (if not zero).
- always attach an estimator of variability, e.g. standard error;
- interpretation?
- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem.
- 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.

Interval Estimation

- Definition (Confidence interval): Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 & \dots & Y_n \end{pmatrix}$ 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
$$\Pr_{\theta}\{L(\boldsymbol{Y}) \leq \theta \leq U(\boldsymbol{Y})\} \geq 1-\alpha,$$
for all $\theta \in \Theta$ with equality for at least one value of θ.
- $1-\alpha$ is called the coverage probability or confidence level.
- Interpretation is more complex.
- 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 contain θ.

Interval Estimation

- Example (The example that says all).
- Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$.
- Then it follows that $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0,1)$ so that

$$\Pr_\mu\{-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96\} = 0.95.$$

- Thus we can deduce

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

- It is clear

$$\Pr_\mu\{\bar{Y} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1.96}{\sqrt{n}}\} = 0.95.$$

- Thus the random interval $[L(\boldsymbol{Y}), U(\boldsymbol{Y})] = [\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}}]$ is a 95% random interval for μ.

Interval Estimation II

- Central Limit Theorem: same argument can yield approximate 95% CI when $Y_1, \dots, Y_n$ are iid, $\mathbb{E}\, Y_i = \mu$ and $\mathbb{V}\text{ar}\{Y_i\} = 1$ regardless of their distribution.
- Notice that the interval is centred at $\bar{Y}$ which is the MLE of μ. Letting the variance take an arbitrary value it is often written:

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

- The length of the interval is $2z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ which depends on σ^2, n and α.
- The parameter σ^2 is outside our control.
- We can however often control n and $1-\alpha$. Increasing n the length of the interval decreases like $1/\sqrt{n}$
- Reducing α or increasing $1-\alpha$ increases the length of the interval, (the dependence is quite non-linear, and 5% is the sweet spot.

Interval Estimation III

- What can we learn from the example we considered?
- Definition (Pivot): A random function $g(\boldsymbol{Y}, \theta)$ is said to be a pivotal quantity or just a pivot if it is a function both of $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ whose distribution does not depend on $\boldsymbol{\theta}$.
- For example $\sqrt{n}\{\bar{Y} - \mu\} \sim \mathcal{N}(0, 1)$ is a pivot in previous example.
- Why is a pivot useful?
- $\forall \alpha \in (0, 1)$ we can determine constants $a < b$ independent of θ such that

$$\Pr_\theta\{a \leq g(\boldsymbol{Y}, \theta) \leq b\} = 1 - \alpha \quad \forall \theta \in \Theta.$$

- If we can manipulate $g(\boldsymbol{Y}, \theta)$ then the above equation yields a CI.

Interval Estimation IV

- Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}(0, \theta)$. The MLE of θ is in this case $\widehat{\theta} = Y_{(n)}$. This has distribution

$$\begin{aligned} \Pr_\theta\{Y_{(n)} \leq x\} &= F_{Y_{(n)}}(x) = \Pr_\theta\left\{\max_i Y_i \leq x\right\} \\ &= \Pr_\theta\{\text{all} \quad Y_i \leq x\} \\ &= \Pr_\theta\{Y_i \leq x\}^n = \left(\frac{x}{\theta}\right)^n. \qquad (1) \end{aligned}$$

This also implies that $T = Y_{(n)}/\theta$ is a pivot as

$$\Pr_\theta\{T \leq t\} = \Pr_\theta\{Y_{(n)}/\theta \leq t\} = \Pr_\theta\{Y_{(n)} \leq t\theta\} = t^n. \qquad (2)$$

- We can now chose a and b such that

$$\Pr_\theta\left\{a \leq Y_{(n)}/\theta \leq b\right\} = 1 - \alpha.$$

- But there are infinitely many such choices. Idea: choose pair $(a; b)$ that minimizes interval's length!

Interval Estimation V

- The solution to this problem is $a = \alpha^{1/n}$ and $b = 1$ which yields

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

- Pivotal quantities can also be used to construct CIs for θ_k when we have a multi-dimensional parameter $\boldsymbol{\theta}$

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

  and the remaining coordinates are also unknown. A pivotal quantity should now be function $g(\boldsymbol{Y}, \theta_k)$ which
- Depends on $\boldsymbol{Y}$ and θ_k but no other parameters;
- Has a distribution independent of any of the parameters (think about the Gaussian problem when the mean is of interest, but the variance is unknown!).

Interval Estimation VI

- Main challenges with pivotal method:
- Hard to find exact pivots in general problems;
- Exact distributions may be intractable.
- Resort to asymptotic approximations...
- In the classical example we would use $a_n\{\widehat{\theta}_n - \theta\} \xrightarrow{\mathcal{L}} \mathcal{N}\{0, \sigma^2(\theta)\}$.

Interval Estimation VII

- What about higher dimensional parameters of interest?
- Definition: (Confidence Region). Let $\boldsymbol{Y}$ be random variables with joint distribution depending on $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. A random subset $R(\boldsymbol{Y})$ of Θ depending on $\boldsymbol{Y}$ is called a $100(1-\alpha)\%$ confidence region for θ if

$$\Pr_\theta\{\boldsymbol{\theta} \in R(\boldsymbol{Y})\} \geq 1-\alpha, \forall \theta \in \Theta,$$

  and equality for at least one value of $\boldsymbol{\theta}$.
- No restriction requiring R to be convex or connected.
- Nevertheless, many notions extend immediately to CR case.