Design of Studies and Comparing Procedures

Surya Tokdar

Uses of frequentist calculations

- So far we have seen the use of frequentist calculations to provide guarantees for statistical procedures.
- There are at least two other major uses of these calculations
- Design of study:
  - Determine how data is to be collected
  - To extract "maximum information" within cost restrictions
- Comparison of procedures:
  - Compare two or more procedures with same guarantees
  - Which makes maximum use of available information?
- The two issues are somewhat overlapping

Sample size determination

- To estimate the prevalence of child malnutrition in a country
- Quantity of interest: p =proportion of malnourished children under age 10
- Survey n children under 10 years, data X = number of malnourished
- What n to use? [Surveying is expensive]

Margin of error consideration

- Model $X \sim Bin(n, p)$, $p \in (0, 1)$
- Want a 95% confidence interval for p with 5% margin of error
- Means interval half-width is no larger than .05
- Use smallest n to meet margin of error

Margin of error calculation

- $100(1-\alpha)\%$ ML interval $\hat{p}_{\text{MLE}} \mp z(\alpha)\sqrt{\frac{\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}{n}}$
- Half width $z(\alpha)\sqrt{\frac{\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}{n}} \le m$ if $n \ge \frac{z(\alpha)^2\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}{m^2}$
- Can calculate the bound if a preliminary estimate of p is available
- Otherwise, use the fact $\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})$ can be at most $1/4$ (for $\hat{p}_{\text{MLE}} = 1/2$), giving the worst case bound:

$$n \ge z(\alpha)^2/(4m^2)$$

- Need at least $n = 385$ for a 95% interval with 5% margin

Power consideration

- Suppose want to test $H_0 : p < 0.25$
- Would be happy to accept H_0 even if $p \in (0.25, 0.33)$
- But care about the difference between $p < 0.25$ and $p > 0.33$
- Size-α test rejects H_0 if $\hat{p}_{\text{MLE}} > 0.25 + z(2\alpha)\sqrt{\frac{\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}{n}}$
- Use minimum n so that power at any $p > 0.33$ is at least b (say 80%)

Power calculation

- Fix $p^* > 0.33$
- Power at p^* equals

$$P_{[X|p^*]}\left(\hat{p}_{\text{MLE}} > 0.25 + z(2\alpha)\sqrt{\frac{\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}{n}}\right)$$
$$= P_{[X|p^*]}\left(T > \hat{\delta}(p^*) + z(2\alpha)\right)$$

where
  - $T = \frac{\hat{p}_{\text{MLE}} - p^*}{\sqrt{\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}/\sqrt{n}} \sim AN(0,1)$
  - $\hat{\delta}(p^*) = \frac{0.25 - p^*}{\sqrt{\hat{p}_{\text{MLE}}(1-\hat{p}_{\text{MLE}})}/\sqrt{n}} \approx \sqrt{n}\frac{0.25-p^*}{\sqrt{p^*(1-p^*)}}$
- So power at p^* is approximately $1 - \Phi(\hat{\delta}(p^*) + z(2\alpha))$

From power to sample size

- Minimum power is at $p^* = 0.33$ (among all $p^* > 0.33$) with $\hat{\delta}(p^*) \approx -0.17\sqrt{n}$
- Need $1 - \Phi(-0.17\sqrt{n} + z(2\alpha)) \geq b$ which happens when

$$n \geq \left\{\frac{z(2\alpha) - \Phi^{-1}(1-b)}{0.17}\right\}^2$$

- For a size-5% test to have at least 80% power at all $p > 0.33$ we need at least $n = 214$.

For linear models

- Similar calculations can be done for linear models: $Y_i = Z_i^T\beta + \sigma\epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} f(0,1)$
- Interval and tests for $\eta = a^T\beta$ are constructed based on $\hat{\eta} = a^T\hat{\beta}_{\text{LS}}$ with property

$$\sqrt{n}(\hat{\eta} - \eta) \sim AN(0, \sigma^2\{a^T M^{-1} a\})$$

where $M = \mathbb{E}[Z_1 Z_1^T]$
- Need a preliminary estimate of σ and M

Wald's two-stage procedure

- Do a pilot with sample size n_0
- Estimate M by $\hat{M}_{n_0} = \frac{1}{n_0}\sum_{i=1}^{n_0} Z_i Z_i^T$
- Estimate σ^2 by $s_{n_0}^2 = \frac{1}{n_0 - p}\sum_{i=1}^{n_0}(Y_i - Z_i^T\hat{\beta}_{n_0})^2$ where $\hat{\beta}_{n_0}$ is the LS estimate based on the pilot
- Calculate required sample size n (to meet a margin m need $n \geq \frac{z(\alpha)^2 a^T \hat{M}^{-1} a}{m^2}$)
- Get another $n - n_0$ many observations
- Interval or test based on $\sqrt{n}(\hat{\eta} - \eta) \sim AN(0, s_{n_0}^2 a^T \hat{M}_{n_0}^{-1} a)$
- Some concerns about whether full sample can be used in the asymptotic variance (read Woodroofe[1])

[1] http://www.stat.lsa.umich.edu/~michaelw/PPRS/1986aam.pdf

Optimal designs for linear models

- Consider data (z_i, Y_i), $i = 1, \cdots, n$, where n has been determined based on cost considerations, and we want to pick $z_1, \cdots, z_n$ from a candidate set $\mathcal{Z}$ so that our intervals are shortest or tests have maximum power
- Because $\hat{\beta}_{\text{LS}} \sim AN(\beta, \sigma^2(Z^TZ)^{-1})$, need $(Z^TZ)^{-1}$ small
- But $(Z^TZ)^{-1}$ is a matrix – need a scalar summary
- Many criteria:
  - *A*-optimality: minimize trace of $(Z^TZ)^{-1}$
  - *D*-optimality: minimize $\det\{(Z^TZ)^{-1}\}$
- Useful when $\mathcal{Z}$ is a discrete set or a bounded set
- Related subject: optimal factorial designs

Comparing procedures

- Between two intervals with same confidence coefficient, we prefer the one with shorter width (need smaller sample size to meet a margin)
- Between two tests of the same size, prefer the one with more power at a detectable distance
- Two considerations:
  1. Optimality (within a regular model, i.e., Gaussian linear model)
  2. Robustness (within a model that encompasses many regular models, i.e., linear models with possibly non-Gaussian errors)
- ML procedures are often optimal but not robust

Most powerful tests

- $X \sim f(x|\theta)$, $\theta \in \Theta$
- Want to test $H_0 : \theta \in \Theta_0$
- A level-α test δ is most powerful at a $\theta_1 \in \Theta \setminus \Theta_0$ if there is no level-α test δ' with $\beta(\theta_1, \delta') > \beta(\theta_1, \delta)$.
- A test δ is said to be the uniformly most powerful level-α test if δ is a most powerful level-α test at every $\theta_1 \in \Theta \setminus \Theta_0$.

The Neyman-Pearson lemma

- Consider $\Theta = \{\theta_0, \theta_1\}$ and we want to test $H_0 : \theta = \theta_0$
- The Neyman-Pearson lemma says that the most power level α test is given by

$$\text{Reject } H_0 \text{ if } \Lambda(X) = \frac{L_X(\theta_1)}{L_X(\theta_0)} > k$$

  where k is such that
  - $P_{[X|\theta_0]}(\Lambda(X) > k) \le \alpha_0$ but
  - $P_{[X|\theta_0]}(\Lambda(X) > k') > \alpha_0$ for all $k' < k$.
  - i.e., k is the smallest threshold satisfying the size condition

UMP test for simple H_0

- Let Θ be a general set. Want to test $H_0 : \theta = \theta_0$
- Suppose there exists a statistic $T(x)$ such that for any $\theta_1 \neq \theta_0$ and any constant $k > 0$ there exists a constant $c > 0$ so that

$$\frac{L_X(\theta_1)}{L_X(\theta_0)} > k \iff T(X) > c$$

  (i.e., for any data x the event on the left happens if and only if that on the right also happens)
- Note that c may depend on both k and θ_1.
- The UMP level-α test rejects H_0 if $T(X) > c(\alpha)$ where $c(\alpha)$ is such that $P_{[X|\theta_0]}(T(X) > c(\alpha)) = \alpha$.
- Why??

UMP test for one-sided hypotheses

- $X \sim f(x|\theta)$, $\theta \in \Theta$, θ is scalar. Test $H_0 : \theta \le \theta_0$.
- Suppose $\{f(x|\theta) : \theta \in \Theta\}$ is a monotone likelihood ratio (MLR) family in a statistic $T(x)$, i.e.,

$$\text{for every } \theta_1 < \theta_2, \frac{f(x|\theta_1)}{f(x|\theta_2)} \text{ is increasing in } T(x)$$

- The UMP level-α test rejects H_0 if $T(X) > c(\alpha)$ where $c(\alpha)$ satisfies $P_{[X|\theta_0]}(T(X) > c(\alpha)) = \alpha$.

UMP unbiased

- Not too many families are MLR in any statistic
- Consequently, not too many UMP tests are known
- A slightly less demanding criterion is to be UMP among all tests that are unbiased, i.e., whose power function at any $\theta \notin \Theta_0$ is at least as much as the size α
- In most examples, ML tests are UMP among this class of tests

Optimal estimation: UMVUE

- In general, tests and intervals are based on estimators
- How do we compare estimators?
- An old favorite of classical statisticians' is the uniformly minimum variance unbiased estimator (UMVUE).
- An estimator $T(x)$ of $\eta = h(\theta)$ is said to be UMVUE if
  1. $T(x)$ is unbiased, i.e., $\mathbb{E}_{[X|\theta]} T(X) = h(\theta)$ $\forall\theta$ and
  2. for any unbiased $\tilde{T}(X)$, $\mathbb{V}\text{ar}_{[X|\theta]} T(X) \le \mathbb{V}\text{ar}_{[X|\theta]} \tilde{T}(X)$, $\forall\theta$

Rao-Blackwell inequality

- Let $T(x)$ be unbiased for $\eta = h(\theta)$
- Suppose $S(x)$ is a sufficient statistic
- The conditional distribution of X given $S(X)$ remains the same for every θ (a consequence of sufficiency)
- Define $T^*(x)$ as $\mathbb{E}\{T(X)|S(X) = S(x)\}$ [conditional expectation taken under any θ, they all give the same]
- Then,
    1. $T^*(x)$ is unbiased and
    2. $\mathbb{V}\mathrm{ar}_{[X|\theta]} T^*(X) \le \mathbb{V}\mathrm{ar}_{[X|\theta]} T(X)$, $\forall\theta$

Information inequality

- There is a limit on how small the variance of an unbiased estimator can be
- A famous result, independently due to Cramér, Rao and Frechét, states the following
- Theorem (Information inequality). If $T(x)$ is unbiased for $\eta = h(\theta)$ then

$$\mathbb{V}\mathrm{ar}_{[X|\theta]} T(X) \ge \dot{h}(\theta)^T [I^F(\theta)]^{-1} \dot{h}(\theta)$$

where $I^F(\theta) = -\mathbb{E}_{[X|\theta]} \frac{\partial^2}{\partial\theta^2} \log f(X|\theta)$

Information inequality, exponential family & UMVUE

- $f(x|\theta) = h(x)e^{\theta' T(x) - A(\theta)}$: a canonical exponential family
- Equivalent parametrization in terms of

$$\eta = \dot{A}(\theta) = \mathbb{E}_{[X|\theta]} T(X)$$

- $T(x)$ is unbiased for η (by definition), $\mathbb{V}\mathrm{ar}_{[X|\theta]} T(X) = \ddot{A}(\theta)$
- Also, $I^F(\theta) = \ddot{A}(\theta)$, so information bound $= \ddot{A}(\theta)$
- So $T(X)$ is UMVUE[2] for η

[2] Does not guarantee $\hat{\theta}_{\text{MLE}}(x) = \dot{A}^{-1}(T(x))$ is UMVUE or even unbiased for θ; a proof that such properties are not too exciting!

Asymptotic relative efficiency

- $X_1, \cdots, X_n \overset{\text{IID}}{\sim} g(x_i|\theta)$, interested in $\eta = h(\theta)$
- Most statistical procedures are built upon estimators $\hat{\eta}$ of η that are not unbiased but satisfy $\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} N(0, se^2(\theta))$.
- Such estimators are (at least) asymptotically unbiased
- The ratio $\frac{se_1^2(\theta)}{se_1^2(\theta)}$ is the asymptotic relative efficiency (ARE) of the estimator $\hat{\eta}_2$ with respect to $\hat{\eta}_1$.
- Efficient estimators give shorter intervals, more powerful tests

Efficiency of MLE

- Usually $\sqrt{n}(\hat{\eta}_{\text{MLE}} - \eta) \xrightarrow{d} N(0, \dot{h}(\theta) T I_1^F(\theta)^{-1} \dot{h}(\theta))$.
- That is, asymptotically, MLE is unbiased and meets the information bound (note, $I^F(\theta) = nI_1^F(\theta)$)
- So information inequality suggests that $\hat{\eta}_{\text{MLE}}$ should have smaller variance than any other estimate that is asymptotically normal with mean η.
- A precise statement is given below.

Efficiency of MLE (contd)

- If $\tilde{\theta}$ is a minimum contrast estimate of θ then $\tilde{\eta} = h(\tilde{\theta})$ satisfies $\sqrt{n}(\tilde{\eta} - \eta) \xrightarrow{d} N(0, \tilde{se}^2(\theta))$ with

$$\tilde{se}^2(\theta) \ge \dot{\mu}(\theta)^T I_1^F(\theta)^{-1} \dot{\mu}(\theta)$$

- That is, the MLE is efficient among all minimum contrast estimators that are asymptotically normal.

Robust procedures

- For non-parametric models, optimal procedures are difficult to find
- A non-parametric model contains many regular parametric sub-models, their corresponding ML procedures are optimal within that sub-model
- However, it might be possible to find a procedure that remains competitive across all sub-models
- Such procedures are called robust (there are other definitions of robustness, but the essence is this)
- Two examples

Median is more robust than mean

- $X_i = \mu + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} f \in \mathcal{F}_{sym}$, $\mu \in (-\infty, \infty)$
- $\mathcal{F}_{sym}$ contains all pdfs that are symmetric around 0
- Two estimators of μ: $\bar{x}$ and x_{med}
- ARE of x_{med} w.r.t $\bar{x}$ is $4f_0(0)\sigma_f^2$, where $\sigma_f^2 = \int x^2 f(x)dx$
- Consider $f = (1-\epsilon)N(0,\sigma^2) + \epsilon N(0, k\sigma^2)$
- ARE $= 64\%$ for $\epsilon = 0$ (best case for mean, being the MLE for that sub-model)
- For $\epsilon > 0$, ARE $\to \infty$ as $k \to \infty$
- Regression counterpart: least-squares vs median (quantile) regression

Wilcoxon's rank-sum test vs t-test

- $X_i \overset{\text{IID}}{\sim} g(x)$, $Y_j \overset{\text{IID}}{\sim} g(x-\theta)$, g arbitrary
- To test $H_0 : \theta = 0$
- Let δ_1 be size-α t-test, δ_2 be size-α W's test
- Fixed desired power $\beta > \alpha$
- Pitman efficiency: Consider a sequence $(\theta_k)_{k\geq 1} \to 0$. Let $n_{k,1}$ and $n_{k,2}$ be sample sizes needed by δ_1 and δ_2 to have power β at θ_k. Relative Pitman efficiency of δ_2 to δ_1 is $\lim_{k\to\infty} \frac{n_{k,1}}{n_{k,2}}$
- Limit does not depend on β
- Relative P efficiency of W's test w.r.t t-test is
  1. 0.95 when g is normal (the best case scenario in favor of t-test)
  2. Can be arbitrarily large (and approach ∞) for non-normal (heavy tailed) g