Amy Braverman

(Jet Propulsion Laboratory)

Basic of Inference
Part 2

Jet Propulsion Laboratory

Keck Institute for Space Studies

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Caltech

Outline

Tools for inference:

- The Central Limit Theorem.
- Confidence intervals.
- Bayesian formalism.
- Summary.

Material on large sample theory based largely on Tom Ferguson's 1996 book, *A Course in Large Sample Theory*, Chapman and Hall.

The Central Limit Theorem (CLT) :

Let $Y_1, Y_2, \ldots, Y_N$ be a sequence of iid random variables, each with expected value $E(Y_n) = \mu_Y$ and variance $var(Y_n) = \sigma_Y^2$, both finite.

Then the distribution of the random variable

$$S_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \frac{(Y_n - \mu_Y)}{\sigma_Y}$$

tends to the standard normal (Gaussian) distribution as $N \to \infty$.

In other words,

$$P(S_N \leq a) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left\{-u^2/2\right\} du \text{ as } N \to \infty.$$

Central Limit Theorem

This can also be written in any of the following ways:

$$\sum_{n=1}^{N} Y_n \xrightarrow{d} Gau(N\mu_Y, N\sigma_Y^2),$$

$$\bar{Y}_N \xrightarrow{d} Gau(\mu_Y, \sigma_Y^2/N),$$

$$\sqrt{N}(\bar{Y}_N - \mu_Y) \xrightarrow{d} Gau(0, \sigma_Y^2),$$

where $\xrightarrow{d}$ indicates convergence in distribution as sample size N gets large. The limiting distribution is called the asymptotic distribution of the statistic.

There is a version of the CLT for independent but non-identically distributed random variables. It is called the Lindeberg-Feller CLT, and it has an extra special condition: that no one term in $var(\sum_{n=1}^{N} Y_n)$ dominates in the limit.

Central Limit Theorem

CLT for random vectors:

Let $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_N$ be a sequence of iid random vectors, each with expected value $E(\mathbf{V}_n) = \boldsymbol{\mu}_\mathbf{V}$ and variance $var(\mathbf{V}_n) = \boldsymbol{\Sigma}_\mathbf{V}$. Then,

$$\sqrt{N}(\bar{\mathbf{V}}_N - \boldsymbol{\mu}_\mathbf{V}) \stackrel{d}{\to} Gau(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{V}).$$

CLT for functions of random vectors (Cramér's Theorem):

Suppose $\mathbf{g}(\cdot)$ is a vector-valued function with continuous derivative $\dot{\mathbf{g}}(\mathbf{v})$. Then,

$$\sqrt{N}\left(\mathbf{g}(\mathbf{V}_N) - \mathbf{g}(\boldsymbol{\mu}_\mathbf{V})\right) \stackrel{d}{\to} Gau\left(\mathbf{0}, \mathbf{g}(\boldsymbol{\mu}_\mathbf{V})\, \boldsymbol{\Sigma}_\mathbf{V}\, \mathbf{g}(\boldsymbol{\mu}_\mathbf{V})^T\right).$$

Central Limit Theorem

The CLT and Cramér's Theorem are *extremely* useful because many estimators end up being functions of sums (or averages) of iid random variables/vectors.

- Sample variance:

$$S_N^2 = N^{-1} \sum_{n=1}^{N} (Y_n - \bar{Y}_N)^2, \quad \sqrt{N}\left(S_N^2 - \sigma_Y^2\right) \xrightarrow{d} Gau(0, \mu_{Y4} - \sigma_Y^4),$$

where $\mu_{Y4} = E(Y_n - \mu_Y)^4$.

Note that S_N^2 can be thought of as a function of the (single realization of the) sample, $\mathbf{Y}$.

Central Limit Theorem

- Sample correlation coefficient for the bivariate random vectors, $\mathbf{V}_n = (V_{1n}, V_{2n})^T$:

$$r_N = \frac{S_{12N}}{\sqrt{S_{11N} S_{22N}}}, \quad S_{ijN} = \frac{1}{N} \sum_{n=1}^{N} (V_{1n} - \bar{V}_{1N})(V_{2n} - \bar{V}_{2N}),$$

$$\rho = \frac{E(V_{1n} - \mu_{V_1})(V_{2n} - \mu_{V_2})}{\sqrt{E(V_{1n} - \mu_{V_1})^2 E(V_{2n} - \mu_{V_2})^2}},$$

$$\sqrt{N}(r_N - \rho) \xrightarrow{d} Gau(0, \gamma^2),$$

where γ^2 is an ugly expression involving true variances and covariances.

The point is, we know what it is.

Other statistics for which the CLT holds:

- Sample quantiles (median, quartiles, etc.)
- Rank (order) statistics
- Chi-squared statistics
- Extrema
- Many others

Central Limit Theorem

CLT for dependent sequences of random variables:

- m-dependence: $Y_1, \ldots, Y_s$ and $Y_{m+s+1}, Y_{m+s+2}, \ldots$ are independent for any choice of s (independence of sets separated by m).
- Stationary: the joint distribution of $(Y_t, \ldots, Y_{t+s})$ does not depend on t (joint distribution same everywhere).

Central Limit Theorem

CLT for dependent sequences of random variables:

If $Y_1, Y_2, \ldots, Y_N$ is a stationary, m-dependent sequence then

$$E\left(\sum_{n=1}^{N} Y_n\right) = N\mu_Y, \quad var\left(\sum_{n=1}^{N} Y_n\right) = \sum_{n_1=1}^{N}\sum_{n_2=1}^{N} cov(Y_{n_1}, Y_{n_2}),$$
$$= N\,var(Y_n) + 2(N-1)\,cov(Y_n, Y_{n+1}) + 2(N-2)\,cov(Y_n, Y_{n+2}) + \ldots + 2(N-m)\,cov(Y_n, Y_{n+m}) \quad \text{for } N \geq m,$$
$$\equiv \tau^2,$$

and

$$\sqrt{N}(\bar{Y}_N - \mu_Y) \xrightarrow{d} Gau(0, \tau^2).$$

Central Limit Theorem

CLT for general MLE:

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} Gau(0, \mathcal{I}(\theta)^{-1}),$$

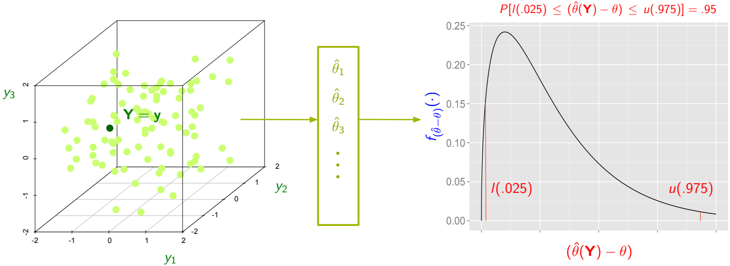
where $\hat{\theta} = \hat{\theta}(\mathbf{Y})$ is a function of the sample, and $\mathcal{I}(\theta)$ is the Fisher Information in random vector **Y** about θ.

$$\psi(\mathbf{y}, \theta) = \frac{\partial}{\partial \theta} \log f_{\mathbf{Y}}(\mathbf{y}, \theta),$$
$$\mathcal{I}(\theta) = var[\psi(\mathbf{Y}, \theta)].$$

We have stated this result for the scalar θ case, and without the slew of required technical conditions.

The catch: have to know $f_{\mathbf{Y}}(\mathbf{y}, \theta)$ in order to compute $\mathcal{I}(\theta)$.

Confidence intervals



- A confidence interval is a random interval computed from a random sample, **Y**, which has a specified probability of containing θ:

$$P(L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})) = .95, \quad \text{with} \quad L(\mathbf{Y}) = \hat{\theta}(\mathbf{Y}) - u(.975), \;\; U(\mathbf{Y}) = \hat{\theta}(\mathbf{Y}) - l(.025).$$

- Example: if $\hat{\theta}(\mathbf{Y}) \sim Gau(0,1)$, $L(\mathbf{Y}) = -1.96$ and $U(\mathbf{Y}) = 1.96$.

Bayesian formalism

Frequentist formalism:

- Everything up to this point treated θ as a fixed but unknown quantity (an ordinary variable).
- Inference based the likelihood function, $L(\mathbf{y}, \theta) = f_{\mathbf{Y}}(\mathbf{y}, \theta)$.

Bayesian formalism:

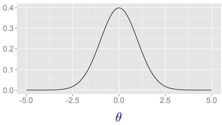
- Treat Θ as a random variable; write the likelihood $L(\mathbf{y}|\theta) = f_{\mathbf{Y}|\Theta}(\mathbf{y}|\theta)$.
- Assert a marginal distribution for Θ: $f_{\Theta}(\theta)$, also sometimes called the “prior" distribution.
- Inference based on the conditional distribution of Θ given $\mathbf{Y}$ (the “posterior"):

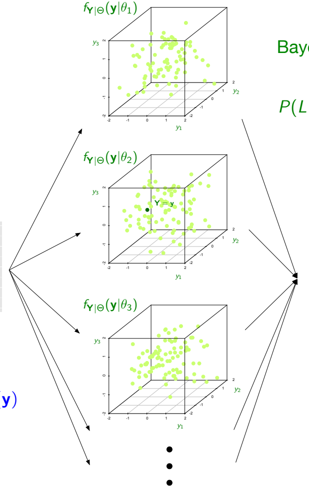
$$f_{\Theta|\mathbf{Y}}(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y}|\Theta}(\mathbf{y}|\theta) f_{\Theta}(\theta)}{f_{\mathbf{Y}}(\mathbf{y})} = \frac{f_{\mathbf{Y}|\Theta}(\mathbf{y}|\theta) f_{\Theta}(\theta)}{\int_{\theta} f_{\mathbf{Y}|\Theta}(\mathbf{y}|\theta) f_{\Theta}(\theta)\, d\theta} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} = P(A|B).$$

Bayesian formalism

$$\Theta \sim f_\Theta(\theta),$$
$$\mathbf{Y} \sim f_{\mathbf{Y}|\Theta}(\mathbf{y}|\theta),$$
$$\Theta|\mathbf{Y} \sim f_{\Theta|\mathbf{Y}}(\theta|\mathbf{y})$$

$f_\Theta(\theta)$



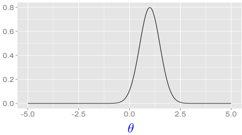


Use sufficient statistic $\hat{\theta} = g(\mathbf{y})$ in place of $\mathbf{y}$ if one exists.

Bayesian confidence interval:

$$P(L \le \Theta \le U|\mathbf{Y} = \mathbf{y}) = \int_L^U f_{\Theta|\mathbf{Y}}(\theta|\mathbf{y})\, d\theta,$$
$$L = F^{-1}_{\Theta|\mathbf{Y}}(.025), \quad U = F^{-1}_{\Theta|\mathbf{Y}}(.975)$$

$f_{\Theta|\mathbf{Y}}(\theta|\mathbf{y})$



$$f_{\Theta|\mathbf{Y}}(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y}|\Theta}(\mathbf{y}|\theta) f_\Theta(\theta)}{f_{\mathbf{Y}}(\mathbf{y})}$$

Summary

Comments:

- It all boils down to how you want to model the unknown parameter: random variable or not.
- Give Θ a flat (uniform or otherwise “non-informative") prior and you get the same answer as you would get from the Frequentist likelihood.
- My opinion: Bayesian formalism is more complete, more flexible, and lends itself to conditional modeling. Easier to use for scientific applications.
- Finally, whether you are a Frequentist or a Bayesian, you still have to know or assume things about the distributions involved in order to use these analytical solutions.

Bibliography

- *Mathematical Statistics and Data Analysis* by John Rice, Wadsworth, 1995.
- *Statistical Inference* by George Casella and Roger L. Burger, Wadsworth, 1990.
- *A Course in Large Sample Theory* by Thomas S. Ferguson, Chapman and Hall, 1996.

Next

The CLT works for many well-behaved statistics, but what about those that are not based on sums? In the next module, we look at resampling procedures which can be very useful in such situations.