The basics of statistical inference

Susan Thomas
IGIDR, Bombay

August 14, 2008

Recap

- The model is estimated by applying the likelihood approach.
- **Estimators** are functions of the data.
  Estimators are also called **statistics**.
- **Estimates** are the values of the estimator for a given set of data.
- Estimates are random variables, with some distribution.
- Inference is the link between the estimate (from a sample) to the population parameter.
- Inference is based on statistical theory of large numbers.

Key results in asymptotic theory

- Two useful results in probability that form the statistical base of econometric inference:
  1. Law of large numbers.
  2. Central limit theorem.

Theorem #1: Law of Large Numbers

Let $(Y_1, Y_2, \ldots, Y_N)$ be random variables that are independent and identically distributed.
Let the distribution have expectation $E(Y) = \theta$.
Then, if $\bar{Y}$ is the sample average, calculated as:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

As $N \rightarrow \infty$,

$$P(|\bar{Y} - \theta| < \delta) \rightarrow 1, \qquad \forall \delta > 0$$

We say that $\bar{Y}$ converges in probability to θ.

Interpreting the LLN

- We assume that all the observations are drawn from exactly the same distribution.
- LLN says that
  the difference between the sample mean ($\bar{Y}$) and the population mean (θ)
  keeps shrinking (or becomes less than δ, which we take as a very small number, $\delta = 0.0001$).
  as the sample size gets larger ($n \rightarrow \infty$)
  with probability one.

Theorem #2: Central Limit Theorem

Let $(Y_1, Y_2, \ldots, Y_n)$ be random variables that are independent and identically distributed.
The distribution is assumed to have expected value θ and **finite variance,** σ^2
If the sample average is $\bar{Y}$ calculated as:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

As $n \to \infty$,

$$P\left(\frac{\bar{Y} - \theta}{\sqrt{\sigma^2/n}} \leq x\right) \to P(X \leq x), \qquad \forall x \in R$$
$$\text{where } X \sim N[0, 1]$$

We say that $\bar{Y}$ is asymptotically distributed as $N[\theta, \sigma^2/n]$.

Interpreting the CLT

- We assume that all the observations are drawn from exactly the same distribution.
- The CLT holds true for the sum of a set of rvs.
  The distribution of the rv can be one out of a broad range of distributions. The rv does **not** have to the gaussian distributed.
- Even though the CLT states it is true as $n \to \infty$, it works well even in finite samples for a symmetric distribution.

Sampling distributions to population distribution

- Using the LLN and CLT, we can get the sampling distribution for any sample statistic/estimate.
  For example, consider the *mean* estimator of a univariate distribution, $\mu = E(x) = \sum_i x_i/n$.
- Generically, the sampling distribution for the sample mean can be derived as:
  If $(x_1, \ldots, x_n)$ *are a sample of rv from a population with constant mean* μ *and variance* σ^2*, then the sample mean* $\bar{x}$ *is rv with a distribution with mean* μ *and variance* σ^2/n*.*
- Further, we set a restriction on the variance being finite using the CLT and get: *If* $(x_1, \ldots, x_n)$ *are a sample of rv from a population with constant mean* μ *and finite variance* σ^2*: then the sample mean* $\bar{x}$ *is rv, distributed as gaussian with mean* μ *and variance* σ^2/n*.*

Distribution for estimates based on a Bernoulli rv

Recap: important parameters of a distribution

- Mean or expected value: $E(x) = \mu$
- Variance: $E[(x - \mu)^2] = \sigma^2$
  Standard deviation: σ
- Skewness: $E[(x - \mu)^3]$
- Kurtosis: $E[(x - \mu)^4]$

E($\hat{\theta}$) of a Bernoulli rv sample

- For a Bernoulli rv, y, $E(y) = \theta$.
  The sample mean, $\hat{\theta} = E(\bar{y})$. What is it's distribution?
- Expected value of $\hat{\theta}$:

$$
\begin{aligned}
E(\hat{\theta}) = E(\bar{y}) &= E(\frac{1}{n}\sum_{i=1}^{n} Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \theta = \theta
\end{aligned}
$$

Var($\hat{\theta}$) of a Bernoulli rv sample

- Variance of $\hat{\theta} = \mathrm{var}(\hat{\theta})$

$$\mathrm{var}(\hat{\theta}) = \frac{1}{n^2}\,\mathrm{var}\left(\sum_{i=1}^{n} Y_i\right)$$

Because they are iid

$$\mathrm{var}(\hat{\theta}) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{var}(Y_1)\right) = \frac{\theta(1-\theta)}{n}$$

The standard deviation of $\hat{\theta}$ is called the *standard error* of the estimator.

Interpreting $E(\hat{y})$ for a Bernoulli rv

- $E(\hat{y}) = \theta$
  The expected value of the sample mean is the population mean.
- This is irrespective of the value of θ. We say that $\hat{\theta}$ is an *unbiased* estimator of θ.
  (Note: We didn't asymptotic theory to make this statement.)

Interpreting $E(\hat{y})^2$ for a Bernoulli rv

- The standard error of $\hat{\theta}$ is:

$$\text{se}(\hat{\theta}) = \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}$$

  Higher *n*, the lower the statistical uncertainty of $\hat{\theta}$ around θ.

- Using **Chebyshev's inequality**, given a rv $\hat{\theta}$ and a positive constant σ, we can say:

$$P(\mu - \mathrm{k}\sigma \leq \hat{\theta} \leq \mu + \mathrm{k}\sigma) \geq 1 - \frac{1}{k^2}$$

  (Note: We can choose any *k* such that the range of values for $\hat{\theta}$ will fall between $\mu + k\sigma$ and $\mu - k\sigma$.)

Interpreting $E(\hat{y})^2$ for a Bernoulli rv

- Using CLT, we refine this:
- We create a standardised form of $z = \hat{\theta}$ as:

$$z = (\hat{\theta})/\mathrm{se}(\theta) = (\hat{\theta})/(\sigma^2/n)$$

- Then:

$$\begin{aligned} z &= \sqrt{(\hat{\theta} - \theta)/((\theta(1-\theta)/n)} \\ &= n\sqrt{\hat{\theta} - \theta}/\sqrt{(\theta(1-\theta))} \\ &\sim N(0,1) \end{aligned}$$

- Now we know that setting $k = 2$, we cover a little more than 95% of probable θ values. Or,

$$P\left[\left(\theta - 2\sqrt{\frac{\theta(1-\theta)}{n}}\right) \leq \hat{\theta} \leq \left(\theta + 2\sqrt{\frac{\theta(1-\theta)}{n}}\right)\right] \approx 95\%$$

Interval based inference about girl vs. boy birth probabilities: the UK dataset

- $E(y) = \hat{\mu}_y = \sum_{i=1}^{N} y_i = 0.4876$
- Variance of y = $E(y - \hat{\mu}_y)^2 = E(y)^2 - E(\hat{\mu}_y)^2 = 0.25$
- Is this statistically different from $\theta = 0.5$?
- From the CLT, we know that the sampling distribution of the $\hat{\theta}$ estimate is a normal distribution.
  Using this, we calculate that with 95% confidence, the range of θ can be derived from this sample as:

$$0.4862 \leq \theta \leq 0.4886$$

- $\theta = 0.5$ does **not** fall in this range.
- Therefore, it appears unlikely that the probability of a girl child being born is the same as the probability of a boy child being born.

Things to remember about interval estimates

- The interval is a rv: the range values change from sample to sample.
- The inference statements says:
  Across repeated samples, there is a “confidence level”% that the interval will contain the population parameter.
- However, there is no direct link between confidence intervals and probability theory.
  Thus, inference falls back upon statistical tests like the LR test.

Hypothesis testing for estimation statistics

The distribution for the LR-test statistic

- The LR test statistic is calculated as

$$\text{LR} = -2 \log \left(L_{\text{restricted}} / L_{\text{unrestricted}} \right)$$

- It can be shown that the LR test has the same distribution as a standardised normal variable:

$$\left[\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\theta(1 - \theta)}} \right]^2$$

- The above has the form of a squared standard normal rv.
- The distribution of a squared standard normal rv is a "χ^2" distribution with one degree of freedom.
- Therefore, the calculated value of the LR test statistic can be compared with a "critical value" of the $\chi^2(1)$ distribution.

Features of the $\chi^2(n)$ distribution

- If $x \sim N(0,1)$, then $z = x^2 \sim \chi^2(1)$.
- First and second moments of a $\chi^2(1)$ distribution:
  - E(z) = 1
  - E($(z - E(z)^2$) = 2
- If $y = \sum_i^n x_i$ and $x_i \sim N(0,1)$, and x_i are independent draws, then $y \sim \chi^2(n)$.
- First and second moments of a $\chi^2(n)$ distribution:
  - E(y) = n
  - E($(y - E(y)^2$) = 2n
- If $x_1 \sim \chi^2(n_1)$ and $x_2 \sim \chi^2(n_2)$, x_1, x_2 are independent, then

$$y = x_1 + x_2 \sim \chi^2(n_1 + n_2)$$

Derivatives of the $\chi^2(n)$ and $N(\mu, \sigma^2)$ distributions

- If $x_1 \sim \chi^2(n_1)$ and $x_2 \sim \chi^2(n_2)$, x_1, x_2 are independent, then
$$y = \frac{x_1/n_1}{x_2/n_2} \sim \mathrm{F}(n_1, n_2)$$
F-distribution has two degrees of freedom, n_1, n_2.
- If $x \sim \chi^2(n)$, and $z \sim N(0,1)$, and x, z are independent, then
$$y = \frac{z}{\sqrt{x/n}} \sim \mathrm{t}(n)$$
t-distribution has one degree of freedom, n.
- Fact: if $x \sim t(n)$, then $x^2 \sim F(1, n)$
- $\chi^2(n), t(n), F(n_1, n_2)$ are all small-sample distributions. As the sample size tends to ∞, each of these converge to other distributions. For example, $t(n) \rightarrow N(0,1)$ as $n \rightarrow \infty$.

Hypothesis testing syntax

- The model or the hypothesis that we start the estimation with is called the *null*. It is denoted as H_0.
  For example, $H_0 = a$ where "a" is a specified value.
- Counter to the null is the *alternative* hypothesis, denoted H_1.
  Sometimes H_1 is explicitly specified as $H_1 = b$. By default, $H_1 = !H_0 \neq a$.
- The test is a procedure which is a function of the sample data, which determines whether to accept H_0 or not.
  For example, reject H_0 if the sample statistic is "too far" away from a.
- In the classical approach, the test also splits the sample statistic space into "a rejection" (or "critical") region and an "acceptance" region.
  If the statistic is in the "acceptance" region, H_0 is accepted as true.

Hypothesis testing syntax

- The statistic is based on a random sample, and so is random itself.
  Thus, the same test can give *different results for different samples*.
- Totally, there could be four different kinds of outcomes for the test results vs. the "truth".

	Accept	Reject
Truth	No problem	Type I
False	Type II	No problem

Hypothesis testing syntax

- Out of these, we worry about the errors and try to quantify them:
  - Size of the test: Probability of a *Type I* error.
    Denoted as α. Also called the "significance level" of the test.
  - Power of the test: Probability of that a *Type II* error will *not* happen – the test will reject the null if it is not true.
    Probability of a *Type II* error is denoted as β.
    Power is denoted as $1 - \beta$.
- The econometrician chooses α.
- Type I errors can be eliminated by making the rejection space very small.
  This increases the probability of Type II errors.
- For a given sample, and a given α, we choose a procedure to make β as small as possible.

Applying hypothesis testing to the UK girl vs. boy birth dataset

- The LR statistic is

$$\mathrm{LR} = -2\log\left(L_{\mathrm{restricted}}/L_{\mathrm{unrestricted}}\right)$$

- $H_0 : \theta = 0.5$.
- In our dataset of fraction of girl vs. boy newborns, the likelihood values are:

$$\begin{aligned}\log L_{\mathrm{R}} &= -496290.6\\ \log L_{\mathrm{U}} &= -496033.8\\ \mathrm{LR} &= 513.6\end{aligned}$$

- Does the sample support or reject H_0?

Applying hypothesis testing to the UK girl vs. boy birth dataset

- The sample test statistic is compared against a $\chi^2(1)$ which has the following values at different levels of significance:

$P(\chi^2(1) > x)$	α		
	0.10	0.05	0.01
x	2.706	3.841	6.635

- At a 95% confidence level, the LR-statistic distribution value to **not reject the null** is 3.84 or less.
- The sample gives a value of 513.6.
- This is much larger than the expected $\chi^2(1)$ value.
- We reject the null of equal probability of seeing girls amongst new borns as compared with boys.

Hypothesis testing syntax

- An **unbiased** test: If the power of the test is greater than the size of the test *for all values of parameters*.
- A **consistent** test: If the power of the test becomes 1 as n becomes ∞.
- The **most powerful** test: A test with highest power among the set of all tests with the same aim.
- Most of the time, we try for unbiased and consistent tests. MP tests are difficult to establish.