Fundamentals of Mathematical Statistics

Confidence Intervals

Tyler J. Brough

Department of Data Analytics & Information Systems

UtahStateUniversity

ECN6330 Agenda for 02/04/2011

Brief Review of Some Concepts

Mathematical Statistics

Confidence Intervals

Agenda for Tonight

- Quiz 3
- Brief Review of Statistical Inference
- Discuss further topics in Mathematical Statistics:
    - Large sample properties of estimators
    - Confidence intervals
    - Hypothesis testing

The Normal Distribution

The normal probability density function is defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(X-\mu)^2}{2\sigma^2}, \quad \text{for } -\infty < x < \infty$$

where $E(X) = \mu$ and $Var(X) = \sigma^2$. When a random variable is normally distributed we write $X \sim N(\mu, \sigma^2)$.

The Normal Distribution

A special case is the standard normal distribution, which is defined as follows:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\}, \quad \text{for } -\infty < z < \infty$$

The standard normal cumulative distribution function is denoted by $\Phi(z) = P(Z \leq z)$. Using some basic facts from probability we arrive at the following helpful formulas:

$$P(Z > z) = 1 - \Phi(z)$$
$$P(Z < -z) = P(Z > z)$$
$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$

The Chi-Square Distribution

The chi–square distribution is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1, 2, \ldots, n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the individual Z_i:

$$X = \sum_{i=1}^{n} Z_i^2$$

The new random variable X has a chi–square distribution with n degrees of freedom. This is often written as $X \sim \chi_n^2$.

The Student T Distribution

The t distribution is a workhorse in classical statistics and econometrics. A t distribution is obtained from a standard normal and a chi–square random variable. Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Also assume that Z and X are independent. Then the following random variable

$$T = \frac{Z}{\sqrt{X/n}}$$

has a t distribution with n degrees of freedom. This is denoted by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi–square random variable.

The F Distribution

Another important distribution for statistics and econometrics is the F distribution. To define an F random variable, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{X_1/k_1}{X_2/k_2}$$

has an F distribution with (k_1, k_2) degrees of freedom. We denote this as $F \sim F_{k_1,k_2}$. The order of the degrees of freedom is important. k_1 is the *numerator degrees of freedom* and k_2 is the *denominator degrees of freedom*.

Large Sample Properties of Estimators

We saw with the estimator of μ, $W = Y_1$ that it was an unbiased, but poor estimator. One notable feature of Y_1 is that its variance is the same no matter what its sample size. It is reasonable to require that as the $[n \to \infty, \quad \sigma^2 \to 0]$ sample size increases the estimation procedure improves.
Example: $\bar{Y}$ for estimation of population mean μ

$$s^2 = Var(\bar{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

As $n \to \infty$, $s^2 \to 0$.

Consistency

Let W_n be an estimator of θ based on a sample $Y_1, Y_2, \ldots, Y_n$ of size n. Then W_n is a *consistent estimator* of θ if for every $\epsilon > 0$

$$P(|W_n - \theta| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty$$

If W_n is not consistent for θ, we say it is *inconsistent*. When W_n is consistent, we also say that θ is the probability limit of W_n, written as

$$plim(W_n) = \theta$$

Interpretation:

- The distribution of W_n becomes more and more concentrated about θ, which means that for larger sample sizes, W_n is less and less likely to vary far from θ.

Python Simulation to Show Consistency

- Go to Python, then come back.
- See the code file `lln.py`

The Law of Large Numbers

Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed random variables with mean μ. Then

$$plim(\bar{Y}) = \mu$$

Interpretation:

- If we want to estimate the population average μ, we can get arbitrarily close to μ by choosing a sufficiently large sample.

Properties of the Probability Limit

Propery PLIM1:
Let θ be a parameter and define a new parameter $\gamma = g(\theta)$, for some continuous function $g(\theta)$. Suppose that $plim(W_n) = \theta$. Define an estimator γ by $G_n = g(W_n)$. Then $plim(G_n) = \gamma$.
Often stated as: $plim(g(W_n)) = g(plim(W_n))$ for a continuous function $g(\theta)$.
Example: $g(\theta) = a + b\theta$, $g(\theta) = \sigma^2$, $g(\theta) = \frac{1}{\theta}$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp\theta$.

Properties of the Probability Limit

Property PLIM2:
If $plim(T_n) = \alpha$ and $plim(U_n) = \beta$ then

- $plim(T_n + U_n) = \alpha + \beta$
- $plim(T_n U_n) = \alpha\beta$
- $plim\left(\frac{T_n}{U_n}\right) = \frac{\alpha}{\beta}$, for $\beta \neq 0$

An Example

Example: Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample of size n an annual earnings from the population of workers with a high school education with population mean μ_Y.
Let $\{Z_1, Z_2, \ldots, Z_n\}$ be a random sample of size n on annual earnings from the population of workers with a college education with population mean μ_Z.
We wish to estimate the percentage difference in annual earnings between the two groups, which is $\gamma = 100\frac{(\mu_Z - \mu_Y)}{\mu_Y}$, the percentage by which earnings for college grads differs from high school grads.
Because $\bar{Y}$ is consistent for μ_Y, and $\bar{Z}_n$ is consistent for μ_Z it follows that

$$G_n = 100\frac{(\bar{Z}_n - \bar{Y}_n)}{\bar{Y}_n}$$

is a consistent estimator of γ.

Asymptotic Normality

Let $\{Z_n : n = 1, 2, \ldots\}$ be a sequence of random variables, such that for all numbers Z

$$P(Z_n \leq z) \rightarrow \Phi(z) \quad \text{as} \quad n \rightarrow \infty$$

in which $\Phi(z)$ is the standard normal CDF. Z_n is said to have an asymptotic standard normal distribution.
Asymptotic normality holds for large n, we have the approximation $P(Z_n \leq z) \approx \Phi(z)$. This means that probabilities concerning Z_n can be approximated by the standard normal probabilities.

The Central Limit Theorm

Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample with mean μ and variance σ^2. Then

$$Z_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

has an asymptotic standard normal distribution.

Confidence Interval Estimation

A *Confidence Interval* is a range of values, calculated from the sample observations, that are believed, with a particular probability, to contain the true parameter value.

An Example of Confidence Interval Estimation

Example:
Suppose the population has a $N(\mu, \sigma = 1)$ distribution and let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample from this population (assume $\sigma = 1$ known).
The sample average $\bar{Y}$ has a normal distribution with mean μ and variance $\frac{1}{n}$

$$\bar{Y} \sim N\left(\mu, \frac{1}{n}\right)$$

Standardizing $\bar{Y}$ which will give it a standard normal distribution

$$P\left(-1.96 < \frac{\bar{Y} - \mu}{1/\sqrt{n}} < 1.96\right) = 0.95$$

Tells us that the probability that the random interval $[\bar{Y} - 1.96/\sqrt{n}, \bar{Y} + 1.96/\sqrt{n}]$, contains the population mean μ

Confidence Interval Estimation

This allows us to construct an interval estimate of μ
$$
$$
This is called a 95% confidence interval. This is denoted as $\bar{y} \pm 1.96/\sqrt{n}$.

Confidence Interval Estimation

Example: suppose $n = 16$ and $\bar{y} = 7.3$ then

$$7.3 \pm 1.96/\sqrt{16} = 7.3 \pm 0.49 = [6.81, 7.79]$$

Confidence Interval Estimation

Interpretation: the random interval
$[\bar{Y} - 1.96/\sqrt{n}, \quad \bar{Y} + 1.96/\sqrt{n}]$ contains μ with probability 0.95.

In other words, if we sampled from the population and calculated the random interval infinitely many times, it would contain the population parameter μ 95% of the time.

Note: It does not mean:
$P(\bar{Y} - 1.96/\sqrt{n} \leq \mu \leq +1.96/\sqrt{n}) = 0.95$ because parameters θ are *not* variables.

Confidence Interval Estimation

- Python simulation exercise to show confidence intervals.
- See Python code file `interval.py`

The Confidence Interval for the Mean of a Normal Population

Assume $X \sim N(\mu, \sigma)$ and σ is known to be any value, the 95% CI is

Confidence Interval for the Mean of a Normal Population

To allow for unknown σ, we must use an estimate. Let

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2}$$

denote the sample standard deviation. Using s for σ does not preserve the 95% CI because we must use the sample to calculate s. In this case we will rely on the t distribution

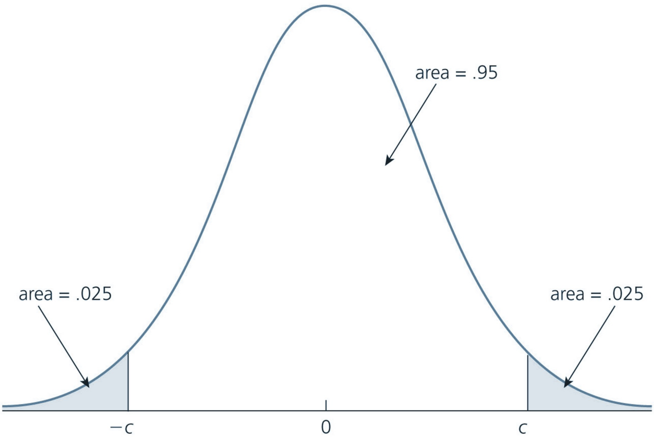
$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Where $\bar{Y}$ is the sample average and s is the sample standard deviation of the random sample $\{Y_1, Y_2, \ldots, Y_n\}$. To construct a 95% CI let c denote the 97.5 percentile in the t_{n-1} distribution.

Confidence Interval for the Mean of a Normal Population

FIGURE C.4

The 97.5th percentile, *c*, in a *t* distribution.



Confidence Interval for the Mean of a Normal Population

In other words:

$$P(-c < t_{n-1} < c) = 0.95$$

For a particular sample

$$[\bar{y} - c\frac{s}{\sqrt{n}}, \quad \bar{y} + c\frac{s}{\sqrt{n}}]$$

Confidence Interval for the Mean of a Normal Population: Example

Example: c is chosen from statistical tables for the t_{n-1} distribution. If $n = 20$ then the $df = 20 - 1 = 19$ Then we have $c = 2.093$ and thus

Example: Consider job training grants on worker production. Let

- $n = 20$
- $c = 2.093$
- $\bar{y} = 1.15$
- $se(\bar{y}) = 0.54$

Confidence Interval for the Mean of a Normal Population: Example

Then we have

$$CI_{0.95} = [-2.28, -0.02]$$

Interpretation: Zero is excluded. We conclude with 95% confidence that average change in scrap rates is not zero.

A Simple Rule of Thumb

A simple rule of thumb for a 95% confidence interval is

- The t distribution approaches the Normal distribution as the degrees of freedom get large.
- In particular for $\alpha = 0.05$, $C_{\frac{\alpha}{2}} \to 1.96$ as $n \to \infty$.
- An approximate 95% CI is

Asymptotic Confidence Intervals for Non-Normal Populations