

Econ 312: Introduction to Econometrics

Review of Probability III

Sang-Yeob Lee

February 16, 2010

Normal distribution

- Symmetric bell shaped density
- Fully characterized by mean and variance $N(\mu, \sigma^2)$
- If $X \sim N(\mu, \sigma^2)$, $\Pr(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) = 0.95$
- Standard normal distribution is the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and is denoted $N(0, 1)$.
- Standard normal C.D.F is often denote by Φ ;

$$\Phi(z) = \Pr(Z < z), \text{ where } z \text{ is a constant}$$

- $\Pr(Z > z) = 1 - \Phi(z)$, $\Pr(Z < -z) = \Pr(Z > z)$ and $\Pr(a < Z < b) = \Phi(b) - \Phi(a)$.

Computing Probabilities Involving Normal Random Variables

- Suppose Y is distributed $N(\mu, \sigma^2)$. Then Y is standardized by computing $Z = \frac{Y-\mu}{\sigma}$
 - ① $\Pr(Y \leq c) = P(Z \leq \frac{c-\mu}{\sigma}) = \Phi(\frac{c-\mu}{\sigma})$
 - ② $\Pr(Y \geq c) = P(Z \geq \frac{c-\mu}{\sigma}) = 1 - \Phi(\frac{c-\mu}{\sigma})$
 - ③ $\Pr(c_1 \leq Y \leq c_2) = P(\frac{c_1-\mu}{\sigma} \leq Z \leq \frac{c_2-\mu}{\sigma}) = \Phi(\frac{c_2-\mu}{\sigma}) - \Phi(\frac{c_1-\mu}{\sigma})$
- For example, suppose Y is distributed $N(1, 4)$, What is the probability that $Y \leq 2$?

$$\Pr(Y \leq 2) = \Pr\left(\frac{Y - 1}{2} \leq \frac{2 - 1}{2}\right) = \Pr\left(Z \leq \frac{1}{2}\right) = \Phi(0.5) = 0.691.$$

Exercise

Compute the following probabilities:

- 1 If Y is distributed $N(1,4)$, find $\Pr(Y \leq 3)$
- 2 If Y is distributed $N(3,9)$, find $\Pr(Y > 0)$
- 3 If Y is distributed $N(50,25)$, find $\Pr(40 \leq Y \leq 52)$
- 4 If Y is distributed $N(5,2)$, find $\Pr(6 \leq Y \leq 8)$

Using STATA to compute the probabilities

- display `normal(z)`: returns the cumulative standard normal distribution. $normal(z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
- display `invnormal(p)` returns the inverse cumulative standard normal distribution: if `normal(z)=p`, then `invnormal(p)=z`.
- display `normal(1.96)` .9750021
- display `invnormal(0.975)` 1.959964

Properties of the Normal Distribution

- 1 If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$
- 2 If X and Y are jointly normally distributed, then they are independent, if, and only if, $Cov(X, Y) = 0$
- 3 Any linear combination of independent, identically distributed normal random variables has a normal distribution.
 - This implies that the average of independent, normally distributed normal variables has a normal distribution.
 - If X_1, X_2, \dots, X_n are independent random variables and each is distributed $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, and therefore $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$

Chi-squared Distribution

- The chi-square distribution is the distribution of the sum of n squared independent standard normal random variables.
- $X = \sum_{i=1}^n Z_i^2$. We write this as $X \sim \chi_n^2$.
- For example, let $Z_1, Z_2,$ and Z_3 be independent standard normal random variables. Then, $X = Z_1^2 + Z_2^2 + Z_3^2$ has a chi-square with 3 degree of freedom.

t-Distribution

- Let $Z \sim N(0, 1)$ and $X \sim \chi_n^2$. Further assume that Z and X are independent. Then, the random variable

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

has a t-distribution with n degrees of freedom (d.f)

- We denote this by $T \sim t_n$
- The t-distribution has a bell shape similar to that of the normal distribution.
- If n (df) is small, it has more mass in the tails than the normal.
- As n goes to infinity, the t-distribution converges to the standard normal distribution.

F-distribution

This will be used for the hypothesis testing in the context of multiple regression analysis.

- Let $X_1 \sim \chi_{k_1}^2$ and $X_2 \sim \chi_{k_2}^2$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = (X_1/k_1)/(X_2/k_2)$$

has a F-distribution with (k_1, k_2) degrees of freedom.

- We denote this as $F \sim F_{k_1, k_2}$

Stata: Cumulative Distributions

- $\Pr(\chi_n^2 \leq x)$: `display chi2(n,x)`
- $\Pr(T_n \geq t)$: `display ttail(n,t)`
- $\Pr(F_{k_1,k_2} \leq f)$: `display F(n1,n2, f)`

Exercise

Compute the following probabilities:

- 1 If Y is distributed t_{15} , find $\Pr(Y \leq 1.75)$
- 2 If Y is distributed t_{90} , find $\Pr(-1.99 \leq Y \leq 1.99)$
- 3 If Y is distributed $N(0,1)$, find $\Pr(-1.99 \leq Y \leq 1.99)$
- 4 If Y is distributed χ_{10}^2 , find $\Pr(Y > 18.31)$
- 5 If Y is distributed $F(7, 4)$, find $\Pr(Y > 4.12)$

Distribution of a sample of data drawn from a population

We will assume simple random sampling

- objects are selected at random from a population and
- each member of population is equally likely to be included in the sample.

Randomness and data

- Prior to sample selection, the value of Y is random because the individual selected is random
- Once the individual is selected and the value of Y is observed, then Y is just a number-not a random
- The data set (Y_1, Y_2, \dots, Y_n) , where $Y_i =$ the i^{th} individual sampled.

i.i.d draws

Because individual #1 and #2 are selected at random, the value of Y_1 has no information about Y_2 . Thus:

- Y_1 and Y_2 are independently distributed.
- Y_1 and Y_2 come from the same distribution, that is Y_1 and Y_2 are identically distributed.
- That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (i.i.d).
- More generally, under simple random sampling, $\{Y_i\}$, $i = 1, \dots, n$ are i.i.d.

This framework allows rigorous statistical inferences about moments of population distribution using a sample of data from population.

Estimation

- \bar{Y} is the natural estimator of the mean, but
 - ① What are the properties of \bar{Y} ?
 - ② Why should we use \bar{Y} rather than other estimator?
 - Y_1 (the first observation)
 - maybe unequal weight-not simple average
 - median
 - ③ The starting point is the sampling distribution of \bar{Y} .

Sampling Distribution of the Sample Average

- The sample average of \bar{Y} , of the n observation Y_1, \dots, Y_n is

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \dots + Y_n) = \frac{1}{n} \sum_{i=1}^n Y_i$$

- \bar{Y} is a random variable, and its properties are determined by the sampling distribution of \bar{Y} .
 - The individuals in the sample are drawn at random. The values of (Y_1, \dots, Y_n) are random.
 - Thus, the function of (Y_1, \dots, Y_n) such as \bar{Y} are random: had a different sample been drawn, they would have taken on a different value.
 - The distribution of \bar{Y} over different sample size n is called the sampling distribution of \bar{Y} .
 - The concept of the sampling distribution underpins all of econometrics.

Things we want to know about the sampling distribution

- What is the mean of \bar{Y} ?
 - If $E(\bar{Y}) = \text{true } \mu$, then \bar{Y} is an unbiased estimator.
- What is the variance of \bar{Y} ?
 - How does $\text{Var}(\bar{Y})$ depend on n ? (remember the famous $1/n$ formula)
- Does \bar{Y} become close to μ as n is large?
 - Law of large number: \bar{Y} is a consistent estimator of μ
- $\bar{Y} - \mu$ appears bell shaped for n large is this generally true?
 - In fact, $\bar{Y} - \mu$ is approximately normally distributed for n large. (Central Limit Theorem)

The mean and variance of the sampling distribution of \bar{Y}

General Case-That is, for Y_i from any distribution.

- mean: $E(\bar{Y}) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \frac{1}{n} \sum_{i=1}^n \mu_Y = \mu_Y$
- variance: $Var(\bar{Y}) = \sigma_Y^2/n$ (Remember this?)
- Implications:
 - 1 \bar{Y} is an unbiased estimator of μ_Y (That is, $E(\bar{Y}) = \mu_Y$)
 - 2 $Var(\bar{Y})$ is inversely proportion to n .
 - the spread of the sampling distribution is proportional to $1/\sqrt{n}$
 - Thus the sampling uncertainty associated with \bar{Y} is proportional to $1/\sqrt{n}$ (larger samples, less uncertainty, but square-root law)

The sampling distribution of \bar{Y} when n is large

For small sample sizes, the distribution of \bar{Y} is complicated, but if n is large, the sampling distribution is simple!

- ① As n increase, the distribution of \bar{Y} becomes more tightly centered to around μ_Y (the Law of Large Number)
- ② Moreover, the distribution of $\bar{Y} - \mu_Y$ become normal (the Central Limit Theorem)

The Law of Large Numbers:

An estimator is consistent if the probability that its falls within an interval of the true population values tends to one as the sample size increases.

If (Y_1, \dots, Y_n) are i.i.d and $\sigma_Y^2 < \infty$, then \bar{Y} is a consistent estimator of μ_Y , that is

$$\Pr(|\bar{Y} - \mu_Y| < \epsilon) \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

which can be written, $\bar{Y} \rightarrow^p \mu_Y$

(“ $\bar{Y} \rightarrow^p \mu_Y$ ” means “ \bar{Y} converges in probability to μ_Y ”.)

(the math: as $n \rightarrow \infty$, $\text{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n} \rightarrow 0$, which implies that $\Pr(|\bar{Y} - \mu_Y| < \epsilon) \rightarrow 1$)

The Central Limit Theorem

If (Y_1, \dots, Y_n) are i.i.d and $\sigma_Y^2 < \infty$, then when n is large the distribution of \bar{Y} is well approximated by a normal distribution.

- \bar{Y} is approximately distribution $N(\mu_Y, \frac{\sigma_Y^2}{n})$ (“normal distribution with mean μ_Y and variance σ_Y^2/n ”)
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0, 1)$ (standard normal)
- That is , “standardized” $\bar{Y} = \frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{Var}(\bar{Y})}} = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$ is approximately distributed as $N(0, 1)$.
- The large is n , the better is the approximation.