Handout #6

Title: Foundations of Econometrics Fall/2015
Course: Econ 367 Instructor: Dr. I-Ming Chiu

Continuous Random Variables and Sampling Distributions (chapter four & six)

What is the main difference between a discrete random variable, say Y, and a continue random variable, X?

Discrete $Y \in Z_+$ (non-negative integers) and its P(Y) is termed probability mass function (PMF).
Continuous $X \in R$ (real numbers) and its P(X) is termed probability density function (PDF).

Probability Density Function (PDF)

$$P(a \le X \le b) = \int_a^b f(x)d(x)$$

The cumulative distribution function F(x) for a continuous random variable X is defined for every number x by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)d(y)$$

Note: $F'(x) = f(x)$

Let p be a number between 0 and 1. The $(100p)^{th}$ percentile of the distribution of a continuous random variable X, denoted by $\eta(p)$, is defined by

$$p = F(\eta(p)) = P(X \le \eta(p)) = \int_{-\infty}^{\eta(p)} f(y)d(y)$$

Note: 1. Median is the 50^{th} (p = 50%) percentile. 2. Recall the IQR = $Q_3 - Q_1$. 3. Quantile regression is often used to understand how features affect the overall distribution of the response.

e.g. 4.7 (pp. 165) Suppose the pdf of the magnitude X of a dynamic load on a bridge (in newtons) is given by

$$f(x) = \frac{1}{8} + \frac{3}{8}x \qquad \text{if } 0 \le X \le 2$$

$$= 0 \qquad \text{otherwise}$$

a) Find the CDF of X
b) $P(1 \le X \le 1.5) = ?$
c) What is the probability that X is at least 1.5?
d) Find the 25th percentile.

Mean: The expected value of a random variable X

Mean (μ) = $E(X) = \int_D x * f(x) d(x)$

Variance: The dispersion of a probability distribution

Variance (X) or (σ_X^2) = $E(X - \mu)^2 = \int_D (x - \mu)^2 * f(x) d(x)$

Standard Deviation (σ) = square root of the variance

$Var(X) = E(X^2) - (E(X))^2 = E(X^2) - \mu^2$

Theorem:

$E(h(X)) = \int_D h(x) * f(x) d(x)$

$E(a + b*X) = a + b*E(X)$, $Var(a + b*X) = b^2*Var(X)$

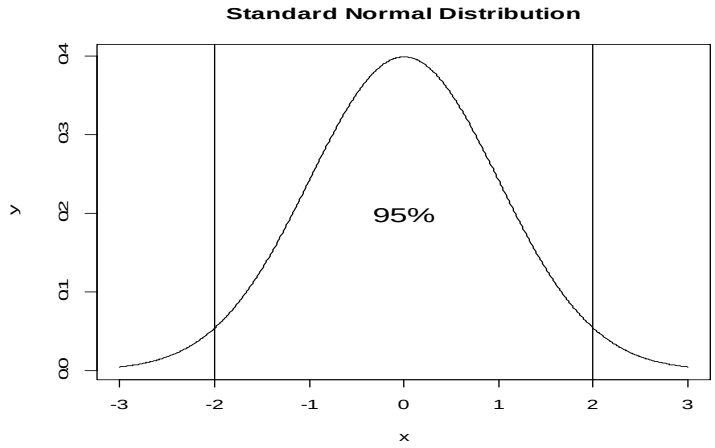
<u>Statistics and Sampling Distribution (Part I)</u>

The Family of Sampling Distribution

(a) Normal distribution (two parameter distribution)
Suppose $X \sim N(\mu, \sigma^2)$; that is random variable X is normally distributed with mean μ and variance σ^2 and has the probability density function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[\frac{-(x-\mu)^2}{2\sigma^2}]$

Standardization $\Rightarrow z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, $f(z) = \frac{1}{\sqrt{2\pi}} \exp[\frac{-z^2}{2}]$;

CDF of Z is $P(Z \leq z)$ and denoted by $\Phi(z)$.



Property of Normal Distribution:

1. $f(\mu + x) = f(\mu - x)$
2. 68-95-99 rule
3. f(x) decreases as x is moving away from μ.

e.g. If $X \sim N(1, 4)$, then what is the probability that X assumes a value no more than 3?

e.g. If $X \sim N(-3, 25)$, then what is the probability that $|X|$ assumes a value greater than 10?

e.g. If $X \sim N(4, 16)$, then what is the probability that X^2 assumes a value less than 36?

Theorem: If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, given that X_1 and X_2 are independent.

Proposition: Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with

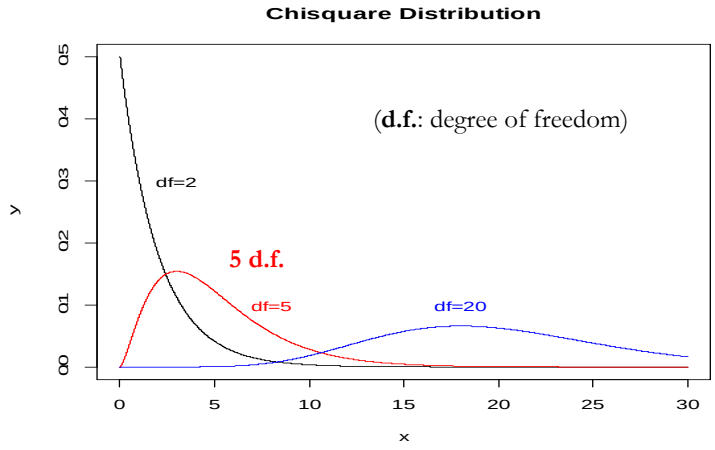
$\mu = np$ and $\sigma = \sqrt{npq}$ (q = 1-p). In particular,

$$P(X \leq x) = B(x; n, p) \cong \Phi(\frac{x + 0.5 - np}{\sqrt{npq}})$$

The above approximation is adequate given that both np and $nq \geq 10$.

e.g 4.26 (pp. 190) Suppose that 25% of all licensed drivers in a state do not have insurance. Let X be the number of uninsured drivers in a random sample of size 50. Find $P(5 \leq X \leq 15)$?

(b) χ^2 Distribution (one parameter distribution; degree of freedom)
If Z_i (i=1…n) are all independently distributed standard normal distribution (i.e., $Z_i \sim N(0, 1)$), then $\sum_{i=1}^{n} Z_i^2$ is said to have a chi-squared distribution with degree of freedom n, χ^2_n.



Theorem: If $Z_1, Z_2, \ldots, Z_n \sim N(0, 1)$ and $Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2$, then $Y \sim \chi^2(n)$

Show that $E(Y) = n$

Proof:

$Z_i \sim N(0, 1) \Rightarrow E(Z_i) = 0$ $Var(Z_i) = 1$

$Var(Z_i) = E(Z_i^2) - (E(Z_i))^2 = 1 \Rightarrow E(Z_i^2) = 1$

$E(Y) = E(\sum_{i=1}^{n} Z_i^2) = 1 + 1 + \ldots + 1 = n$

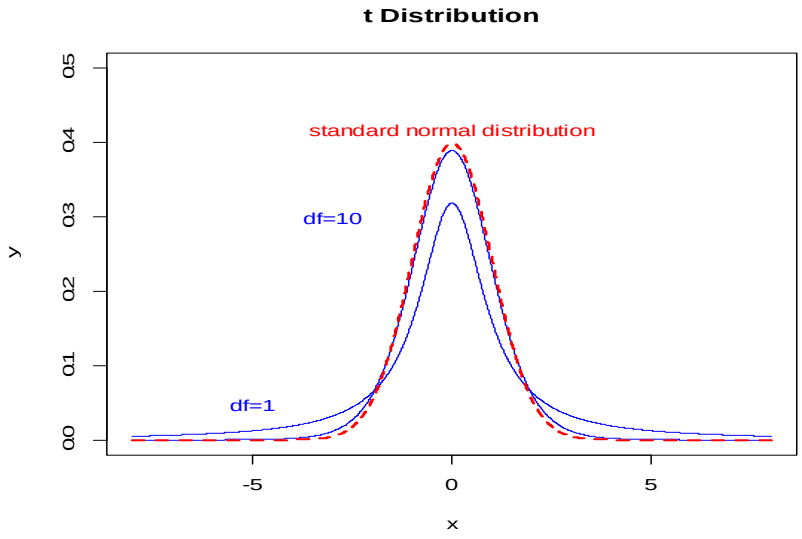
*Note: Var(Y) = 2n; it's easier to show it using Moment Generating function.

Since Y (Chi-squared r.v.) is non-negative, you can define it in $(0, \infty)$.

e.g. Suppose $Y \sim \chi^2(5)$, please find $f(2 < Y < 8)$ using Stata.
e.g. Suppose $Y \sim \chi^2(3)$, please find $\Phi^{-1}(0.95)$ using Stata.

(c) t Distribution (one parameter distribution; degree of freedom)

If Z_i (i=0…n) are all independently distributed standard normal distribution, then $\frac{Z_0}{\sqrt{\sum_{i=1}^{n} Z^2{}_i / n}}$ is said to have a t distribution.



Theorem: $Z \sim N(0, 1)$ and $Y \sim \chi^2(n)$, then $T = \frac{Z}{\sqrt{Y / n}}$ has a t distribution with degree of freedom equals n [using symbol t(n)], given that Z and Y are independent.

Theorem: Let F_n denote the CDF of t(n) and let Φ denote the CDF of N(0, 1). Then

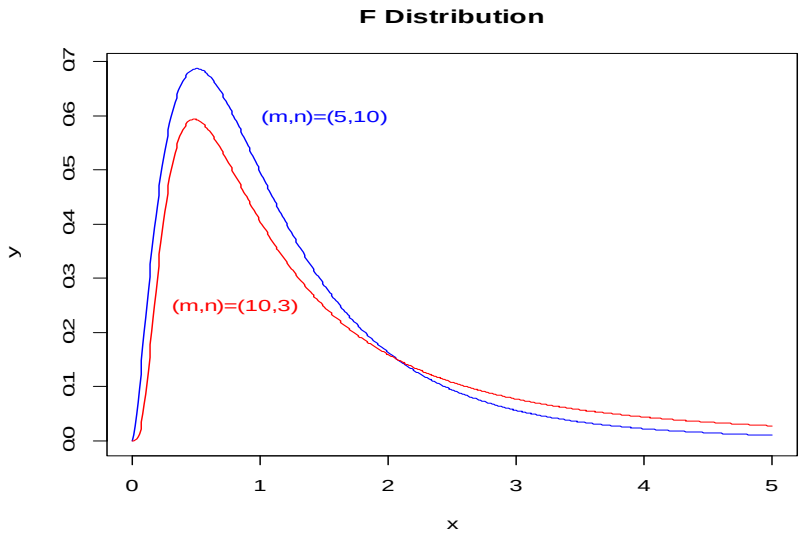
$$\lim_{n\to\infty} F_n(y) = \Phi(y) \text{ for } \forall\, y \in (-\infty, \infty)$$

The above theorem indicates that when degree of freedom in t distribution becomes large, then t distribution can be approximately represented by a standard normal distribution.

e.g. If $T \sim t(14)$ and T assumes a values no greater than -1.5, please find the probability of T using Stata.

Following the above example, let's increase the degree of freedom of n using 100, 500 and 1000.

(d) F Distribution (two parameter distribution; a pair of degree of freedom)
If V_1 and V_2 are two independent random variables having the Chi-Squared distribution with m_1 and m_2 degrees of freedom respectively, then the following quantity follows an F distribution with m numerator degrees of freedom and n denominator degrees of freedom, i.e., (m, n) degrees of freedom.



Theorem: Let $Y_1 \sim \chi^2(m)$ and $Y_2 \sim \chi^2(n)$ be independent variable, then $F = \frac{Y_1 / m}{Y_2 / n}$ is an F distribution with degree of freedom m and n, respectively. It is denoted by F(m, n).

Theorem: If $T \sim t(n)$, then $T^2 \sim F(1, n)$

Proof:

$T = \frac{Z}{\sqrt{\chi^2 / n}} \Rightarrow T^2 = \frac{Z^2}{\chi^2 / n} = \frac{\chi^2 / 1}{\chi^2 / n}$, therefore, $T^2 \sim F(1, n)$

e.g. If $F \sim F(2, 27)$, please find $P(F > 2.5)$ using Stata.

Some important usage of Normal, χ^2, t and F distribution in statistical inference

Definition: A ***statistic*** is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

Suppose a sample of X_i (i=1…n) are drawn from a $N(\mu, \sigma^2)$ RV, then without proof:

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1) \text{ for all i} \quad (1)$$

$$\frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1} \quad (2)$$

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \quad (3)$$

$$\frac{\chi^2_m / m}{\chi^2_n / n} \sim F_{m,n} \quad (4)$$

Where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Theorem: Suppose a random sample of size n is drawn from a population with mean μ and variance σ^2, then the sample average $\overline{X}$ will have the following property:

$E(\overline{X}) = \mu$ & $V(\overline{X}) = \frac{\sigma^2}{n}$, Where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (Can you prove both properties?)

Suppose we also know the population has a normal distribution, then $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Theorem: (Weak) Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be an iid sequence of random variables and $E(X_i) = \mu$, let $S_n = \sum_{i=1}^{n} X_i$.

$\Rightarrow \frac{S_n}{n} \to \mu$, as $n \to \infty$

Alternatively, the above statement can be written as

$\Rightarrow P(\left|\frac{S_n}{n} - \mu\right| < \varepsilon) \to 1$, as $n \to \infty$, $\forall\ \varepsilon > 0$.

Simulation: Often we don't know the features about a population and it makes the inference job challenging. However, we can use simulated data to study the linkage between population features and the corresponding sample characteristics.

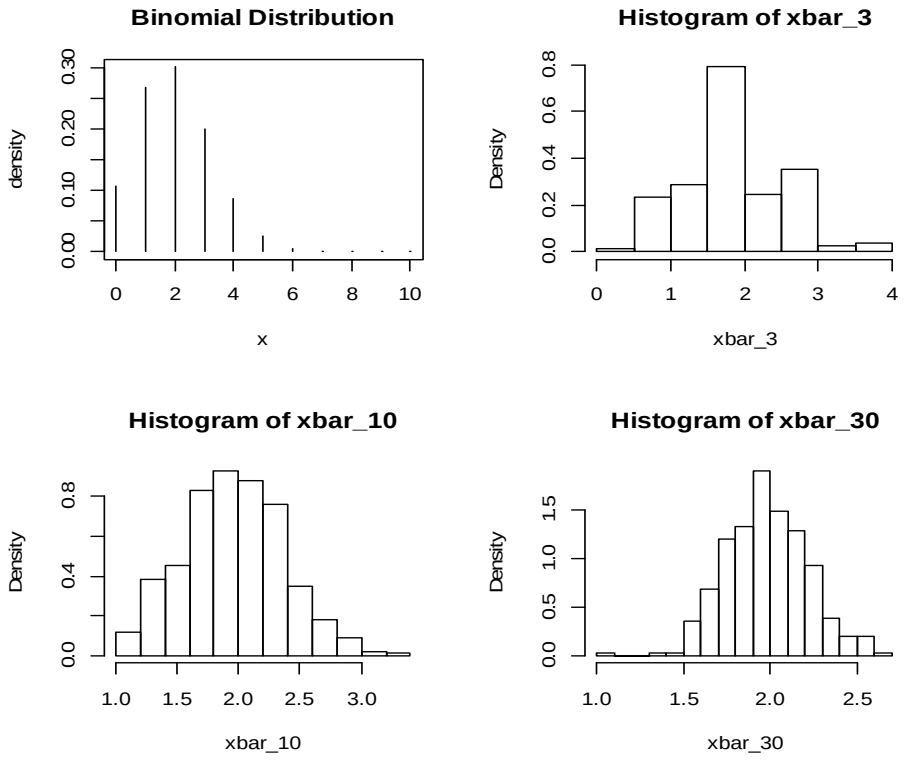
e.g. generate a random sample from a known random variable and study its properties.

The Central Limit Theorem

If large samples (in practice, the size n of each sample is *no less than 30*) are randomly selected from a population with mean μ and variance σ^2, then the sample average $\overline{X}$ will have the following property regardless of the shape of the distribution of the parent population.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Illustrate Central Limit Theorem using Simulation



Explanation: X is a random variable with Binomial distribution (i.e., X~binom(10, 0.2)). The upper left panel shows what its (theoretical) probability distribution looks like; it is asymmetric and skewed right. We choose repeated samples (i.e., 500 samples) with sample size 3, 10 and 30, respectively. The rest of three histograms show the distribution of the sample average (when n =3, 10 and 30), what do you observe?