Massimo Guidolin
Massimo.Guidolin@unibocconi.it
Dept. of Finance

Università Commerciale
Luigi Bocconi

STATISTICS/ECONOMETRICS PREP COURSE – PROF. MASSIMO GUIDOLIN

SECOND PART, LECTURE 1: RANDOM SAMPLING

OVERVIEW

1) Random samples and random sampling
2) Sample statistics and their properties
3) The sample mean: mean, variance, and its distribution
4) Location-scale family and their properties
5) The case of unknown variance: t-Student distribution
6) Properties of the t-Student

RANDOM SAMPLES: "IIDNESS"

- Often, data collected in an experiment consist of several observations on a variable of interest
  - Example: daily stock prices between 1974 and 2013
- In statistics it is often useful to think of such samples as the result of random sampling
- <u>Definition [RANDOM SAMPLING]</u>: The random variables X_1, ..., X_n are called a random sample of size n from the population f(x) if X_1, ..., X_n are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function, f(x)
  - X_1, ... , X_n are called independent and identically distributed random variables with pdf or pmf f(x), IID random variables
  - Pdf = probability density function; pfm = probability mass function (in the case of discrete RVs)
  - Each of the X_1, ... , X_n have the same marginal distribution f(x)

Random Samples: "IIDness"

- The observations are obtained in such a way that the value of one observation has no effect on or relationship with any of the other observations: $X_1, ..., X_n$ are mutually independent
- Because of this property, the joint pdf or pmf of $X_1, ..., X_n$ is:

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

  where $f(x_i; \boldsymbol{\theta})$ is the pdf/pfm and $\boldsymbol{\theta}$ is a vector of parameters that enter the functional expression of the distribution
  - E.g., $f(x_i; \boldsymbol{\theta}) = (1/[2\pi]^{1/2})\exp(-x^2/2)$, the standardized normal distribution
- Soon our problem will be that $\boldsymbol{\theta}$ is unknown and must be estimated
- <u>Example 1</u>: Suppose $f(x_i; \theta) = (1/\theta)\exp(- x_i/\theta)$, an exponential distribution parameterized by θ. Therefore

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}} = \frac{1}{\theta^n} \prod_{i=1}^{n} e^{-\frac{x_i}{\theta}} = \frac{1}{\theta^n} \exp\left[-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right]$$

- While in infinite samples the definition always holds, in finite

SAMPLE STATISTICS

samples, conditions must be imposed—for instance, replacement of draws ("simple random sampling") must be applied

- In finance, most of what we think of, assumes that infinitely-sized samples are obtainable

- When a sample $X_1, \ldots, X_n$ is drawn, some summary of the values is usually computed; any well-defined summary may be expressed as a function $T(X_1, \ldots, X_n)$ whose domain includes the sample space of the random vector $(X_1, \ldots, X_n)$
  - The function T may be real-valued or vector-valued; thus the summary is a random variable (or vector), $Y = T(X_1, \ldots, X_n)$
  - Because the sample $X_1, \ldots, X_n$ has a simple probabilistic structure (because the X_is are IID), the (sampling) distribution of Y is tractable
  - $T(X_1, \ldots, X_n)$ is also called a sample statistic

SAMPLE STATISTICS

- Two important properties of functions of a random sample are:

$$E\left[\sum_{i=1}^{n} g(X_i)\right] = \sum_{i=1}^{n} E[g(X_i)] = \sum_{i=1}^{n} E[g(X_1)] \underset{\text{from identical dstrb.}}{=} nE[g(X_1)]$$

$$Var\left[\sum_{i=1}^{n} g(X_i)\right] = \sum_{i=1}^{n} Var[g(X_i)] + \underbrace{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i\neq j}}^{n} Cov[g(X_i), g(X_j)]}_{=0 \text{ from independence}}$$

$$= \sum_{i=1}^{n} Var[g(X_1)] \underset{\text{from identical dstrb.}}{=} nVar[g(X_1)]$$

- Most of what you think Statistics is, is in fact about sample statistics: the max value of a sample; the minimum value of a sample; the mean of a sample; the median of a sample; the variance of a sample, etc.

• Three statistics provide good summaries of the sample:

$$(\text{Sample mean})\ \bar{X}(X_1, X_2, \ldots, X_n) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Properties of Sample Statistics

$$\text{(Sample variance) } S^2(X_1, X_2, ..., X_n) = \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\text{(Sample std. deviation) } S(X_1, X_2, ..., X_n) = \hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

- <u>Key result 1</u>: Let $X_1,\ldots,X_n$ be a simple random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

$$\boxed{E[\bar{X}_n]} = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu \boxed{= \mu} \text{ (sample mean is unbiased)}$$

$$\boxed{Var[\bar{X}_n]} = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2} n Var[g(X_1)] \boxed{= \frac{\sigma_n^2}{n}}$$

Important to make it unbiased

$$\boxed{E[\hat{\sigma}_n^2]} = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\bar{X}_n^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - 2n\bar{X}_n^2 + n\bar{X}_n^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right]$$

$$= \frac{n}{n-1}\left(E[X_1^2] - E[\bar{X}_n^2]\right) = \frac{n}{n-1}(\sigma_n^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma_n^2}{n} + \mu^2\right) \boxed{= \sigma_n^2} \text{ (sample var. unbiased)}$$

Properties of the Sample Mean

- These are just results concerning moments, what about the distribution of sums of IID samples?
- As $X_1, \ldots, X_n$ are IID, then $Y = (X_1 + X_2 + \ldots + X_n)$ (i.e., the sum variable) has a cdf/cmf that is equal to $P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = P(X_1 \leq x_1)P(X_2 \leq x_2) \ldots P(X_n \leq x_n) = f(x_1)f(x_2) \ldots f(x_n)$
- Thus, a result about the pdf of Y is easily transformed into a result about the pdf of $\bar{X}_n$
- However, this stops here: unless specific assumptions are made about f(X) in the first instance, if n is finite, then we know nothing about the distribution of $\bar{X}_n$
- A similar property holds for moment generating fncts (mgfs)

• Definition [MGF]: The mgf of a random variable X is the transformation: $M_X(s) = E[e^{sX}] = E[\exp(sX)]$

and it is useful for math tractability as $E[X^k] = d^k M_X(s)/dX^k$

PROPERTIES OF THE SAMPLE MEAN

– Because of the assumption of IIDness, then the following holds with reference to the sample mean:

$$
\begin{aligned}
\boxed{M_{\frac{1}{n}[X_1+X_2+...+X_n]}(s)} &= E[\exp(s\frac{1}{n}X_1 + s\frac{1}{n}X_2 + ... + s\frac{1}{n}X_n)] \\
&= E[\exp(s\frac{1}{n}X_1)\exp(s\frac{1}{n}X_2)...\exp(s\frac{1}{n}X_n)] \\
&= E[\exp(s\frac{1}{n}X_1)]E[\exp(s\frac{1}{n}X_2)]...E[\exp(s\frac{1}{n}X_n)] \\
&= \left\{E[\exp(s\frac{1}{n}X_1)]\right\}^n = \boxed{\left\{M_X(\frac{1}{n}s)\right\}^n}
\end{aligned}
$$

- This is fundamental: if you know $M_x(s)$, then you know the MGF of the sample mean. In particular, if

$$M_X(s) = \exp\left(\mu s + \frac{1}{2}\sigma^2 s^2\right) \text{ (mgf of a normal)}$$

then

$$M_{\bar{X}_n}(s) = \left\{M_X(\frac{1}{n}s)\right\}^n = \left\{\exp\left(\mu\frac{s}{n} + \frac{\sigma^2}{2}\frac{s^2}{n^2}\right)\right\}^n = \exp\left(\mu s + \frac{1}{2}\frac{\sigma^2}{n}s^2\right) \Longleftrightarrow \boxed{\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

PROPERTIES OF THE SAMPLE MEAN

- Key result 2: Let $X_1,\ldots,X_n$ be a simple random sample from a normal population with mean μ and variance $\sigma^2 < \infty$, $N(\mu, \sigma^2)$, then

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Result 2 is usefully integrated by two additional properties that are useful in financial econometrics under normality:
  - (i) the sample mean and the sample variance (and S^2_n) are independent;
  - (ii) the $[(n-1)\, S^2_n/\sigma^2]$ transformation of the sample variance has a chi-squared distribution with n-1 degrees of freedom
- The chi-square distribution will play a fundamental role in your studies; its density (for a generic $X \sim \chi^2_p$) is:

Properties of Sample Mean and Variance

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2} \qquad x \in (0, +\infty)\ p \text{ is the number of degrees of freedom}$$

- $\Gamma(\cdot)$ is the gamma function that can be computed recursively
- Two properties of the chi-square are of frequent use:

❶ If Y is a N(0, 1) random variable, then $Y^2 \sim \chi^2_1$, $E[\chi^2_p] = p$, $Var[\chi^2_p] = 2p$

❷ If $X_1, \ldots, X_n$ are independent and $X_i \sim \chi^2_{pi}$ then $X_{p1} + X_{p2} + \ldots X_{pn} \sim \chi^2_{p1+p2+\ldots+pn}$ that is, independent chi squared variables add to a chi-squared variable, and degrees of freedom add up

- These distributional results are just a first step even under the assumption of normality: we have assumed that the variance of the population $X_1, \ldots, X_n$ is known
- In reality: most of the time the variance will be unknown and will have to be estimated jointly with the mean
  - How? Obvious idea, let's try and use S^2

The Case of Unknown Variance

– Here one very old result established by Gosset, who wrote under the pseudonym of "Student", is that

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Longleftrightarrow \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \ \text{ while } \ \frac{\bar{X}_n - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

where t_{n-1} indicates a new, special distribution, the **t-Student with n-1 degrees of freedom**

- This derives from
$$\frac{\bar{X}_n - \mu}{\sqrt{S^2/n}} = \frac{\bar{X}_n - \mu}{\sqrt{S^2/n}}\frac{\sigma}{\sigma} = \frac{\overbrace{(\bar{X}_n - \mu)/(\sigma/\sqrt{n})}^{\sim N(0,1)}}{\underbrace{\sqrt{S^2/\sigma^2}}_{\sim\sqrt{\chi^2_{n-1}/(n-1)}}}$$

  where the distributions at the numerator and denominator are independent and the denominator derives from $[(n\text{-}1)\, S^2_n/\sigma^2] \sim \chi^2_{n-1} \Rightarrow S^2_n/\sigma^2 \sim \chi^2_{n-1}/(n\text{-}1)$

- <u>Definition [t-Student distribution]</u>: Let $X_1,\ldots,X_n$ be a random sample from a $N(\mu,\sigma^2)$ distribution. Then $(\bar{X}_n - \mu)/(S/n^{1/2})$ has a

THE CASE OF UNKNOWN VARIANCE

Student's t distribution with n - 1 degrees of freedom and density

$$f_T(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\sqrt{(n-1)\pi}} \frac{1}{\left(1+\frac{t^2}{n-1}\right)^{\frac{n}{2}}} \qquad t \in (-\infty, +\infty)$$

- Student's t has no mgf because it does not have moments of all orders
- If there are p degrees of freedom, then there are only p- 1 moments: hence, a t_1 has no mean, a t_2 has no variance, etc.
- The problem set makes you check that if T_p is a random variable with a t_p distribution, then $E[T_p] = 0$, if $p > 1$, and $Var[T_p] = p/(p-2)$ if $p > 2$

- One exercise in your problem set, also derives another useful characterization
- Key result 4: If $T \sim t_p$, then $\lim_{p\to\infty} f(t;p) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ or $T \xrightarrow{D} N(0,1)$

  In words, **when p→∞, a t-Student becomes a standard normal distribution**

Useful Notions Reviewed in This Lecture

Let me give you a list to follow up to:

- What is a random sample and what it means to be IID
- What is a sample statistic and how it maps into useful objects in finance and economics
- Sample means, variances, and standard deviations and their properties
- The moment generating function
- The chi-square distribution and its moments
- The t-Student distribution and its properties
- Relationship between t-Student and normal distribution