# Basic Statistics

Prof. Sergio Focardi PhD
Email: sergio.focardi@edhec.edu

Suggested reference text with many examples:

Svetlozar T. Rachev, Markus Hoechstoetter, Frank J. Fabozzi CFA, Sergio M. Focardi, *Probability and Statistics for Finance*, Wiley, August 2010

# Descriptive Statistics

- Introduction
- Financial Data and Basic Financial Concepts
- Graphical Data Analysis
- Numerical Data Analysis

# The Nature of Statistics

- Methods to collect and analyze data…
- Descriptive statistics: how to summarize data to capture main features of large data sets
- Statistical inference: making judgments, estimates, forecasts about a larger group
- From an observed smaller group
- Foundation for statistical inference is probability theory

# Populations and Samples

- An important distinction:
- The larger group versus the smaller group
- Population: All members of a specified group
- Any quantitative measure of a population characteristic is called a parameter, e.g. the mean
- Often too costly to observe the entire population (surveys).
- Sample: a subset of a population
- The whole idea of sampling is to take samples that represent well the population
- A sample statistic is a quantitative measure of a sample.
- Statistical inference consists in estimating an unknown population parameter using a sample statistic, e.g. the average

# SECTION 1:

# FINANCIAL DATA AND BASIC FINANCIAL CONCEPTS

# Financial Data and UHFD

- Based on trading
- In asset management and risk management we deal with prices and returns
- Price discovery process is the process of discovery at what price a transaction can be performed
- In modern electronic exchanges a Tick is the set of information about a trade identified by a Time Stamp
- Tick by Tick Data (trade price, trade time, and volume traded)
- http://www.tickdatamarket.com/
- Generally called Ultra High Frequency Data UHFD
- Irregular spacing

# Frequency of Ticks

- Number of ticks per day vary widely
- For example, a study of High Frequency Trading activities relative to 120 stocks traded on the NYSE by Brogaard (2010)
- Found trading frequencies ranging from 8 transactions per day for the least traded stocks
- To 60,000 transactions per day, roughly 2 transactions per second on average, for the most traded stocks.

# HFD

- UHFD may be transformed into equally spaced data (1-minute, 5-minute intervals)
- Available for stocks, futures, foreign exchange…
- Called generically high-frequency or intraday data - HFD
- Available for quantitative analysis for the last ten years
- Created new ways of computing statistics such as volatility

# Daily data

- Data with daily frequency
- Might be recorded at the end of the day or in other moments
- Some financial data are not available daily, typically because related to variables that move too slowly in any single day
- Examples of slow-moving data include Book to Price Ratio, Earnings Per Share, etc…

# Bid-ask spread

- A fundamental feature of price data is the bid-ask spread
- The bid-ask spread is the difference between the price at which a security is quoted for immediate sale and the price at which it is quoted for immediate purchase
- The bid-ask spread is a measure of liquidity
- A small bid-ask spread indicates that it is possible to sell at current market prices
- A large bid-ask spread indicates that a trader might need to weight a long period before completing the desired transaction with possibly added cost

# Data types

- Discrete (minimum price change rules) vs. continuous data which can assume any value
- Univariate vs. multivariate data
- Intertemporal (time series of a stock price) vs. cross-sectional data
- Indices and portfolios (equally-weighted, value weighted)
- Grouped data
- Data Analysis (graphical and numerical) depends on form and type of data

# Holding Period Return

- Definition of returns:

$$R_t = \frac{P_t - P_{t-1} + D_t}{P_{t-1}}$$

where

$P_t$ = price per share at the end of time t

$P_{t-1}$ = price per share at the end of time t−1

$D_t$ = cash distributions received during time period t

- This formula can be used to compute the holding period return of any asset over a minute, day, week, month, year, a return is associated with a time interval (daily or monthly returns…)

- Return has no unit of currency attached to it but beware of expressing returns in a different currency than home currency because the exchange rate can change during the holding period

# Prices and returns

- We can reconstruct prices from returns and initial prices using the formulas for compounded returns

$$R(t)+1=\frac{P(t)}{P(t-1)}$$

$$P(t)=P(0)\times(R(1)+1)\times\cdots\times(R(t)+1)$$

- If compounded returns are constant prices grow exponentially:

$$P(t)=P(0)\times(R+1)^{t}$$

# Logprices and logreturns

- To avoid the non-linearities inherent in the concept of compounded return we often use logprices and logreturns
- Logprices are the logarithms of prices
- Logreturns are the differences of the logprices
- The following holds:

$$p(t) = \log(P(t))$$
$$r(t) = p(t) - p(t-1)$$
$$r(t) = \log(P(t)) - \log(P(t-1)) = \log\left(\frac{P(t)}{P(t-1)}\right) = \log(R(t)+1)$$

- If $R(t) << 1$ then $r(t) = \log(R(t)+1) \approx R(t)$

# Where you will find these concepts

- Financial data are obviously fundamental for all type of quantitative financial analysis
- Time series data on prices and returns are used in in asset and risk management as well as in corporate finance
- The use of logprices is typical of time series analysis which is used in asset and risk management

# SECTION 2:
# GRAPHICAL DATA ANALYSIS

# Frequency distribution

- Example: Monthly returns on the NYSE Composite Index
- Frequency table

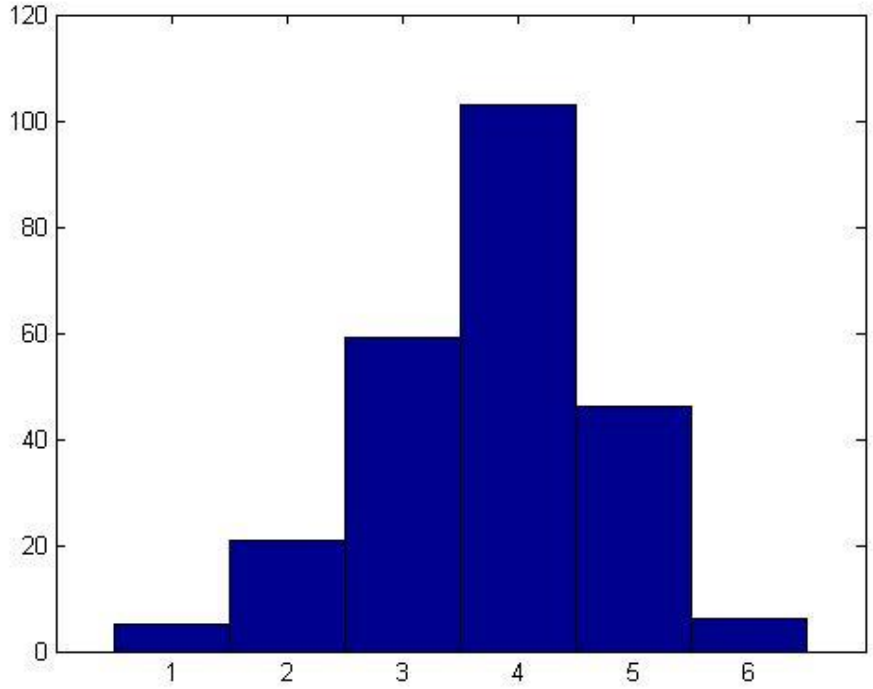
| Return Interval | Frequency |
|---|---|
| X≤ −8% 5 | 5 |
| −8% ≤ x≤ −4% | 21 |
| −4% ≤ x ≤0% | 59 |
| 0% ≤ x ≤ 4% | 106 |
| 4% ≤ x ≤ 8% | 43 |
| x > 8% | 6 |

# Construction of frequency tables

- Sort the data in ascending order
- Calculate the range of data, Range= Max value minus min value
- Decide on the number of intervals in the frequency distribution, k
- Determine interval width as Range/k.
- Determine the intervals by successively adding the interval width to the minimum value to determine the ending points of intervals, stopping after reaching an interval that includes the maximum value.
- Count the number of observations falling in each interval.
- Construct a table of the intervals listed from smallest to largest that shows the number of observations falling in each interval.

# Histogram

- NYSE Index; the height of each «bin» is equal to the number of times the returns of the index fall into the ith bin



# Relative and cumulative frequency

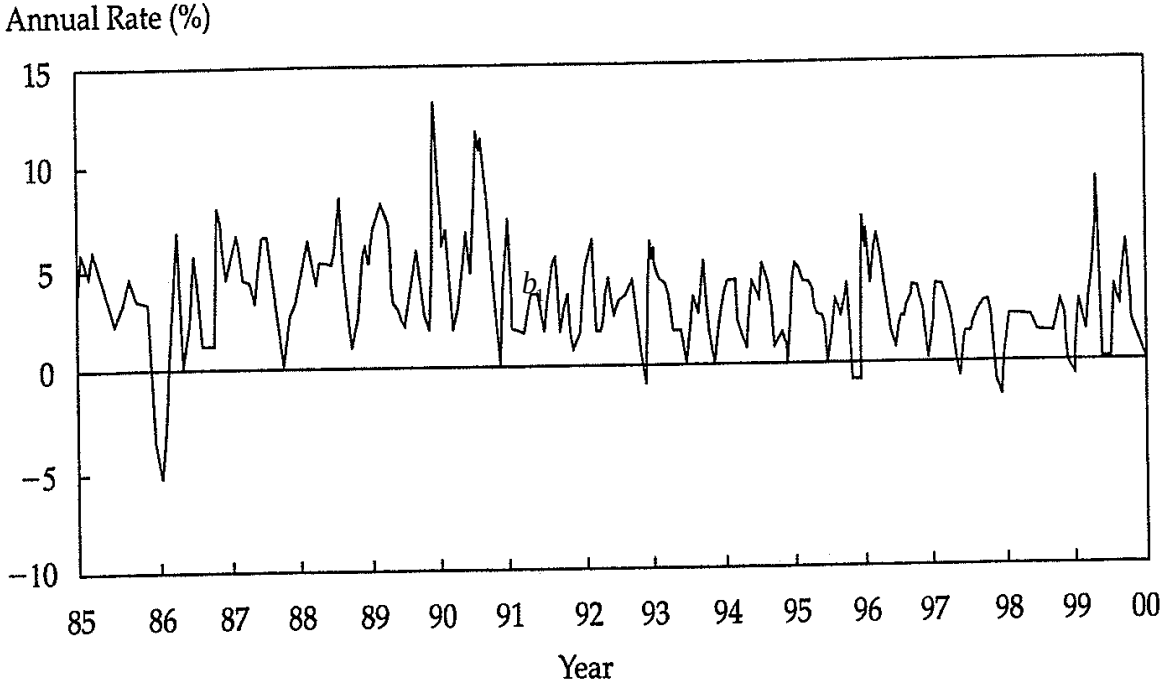
Definition:

- The relative frequency is the absolute frequency of each interval divided by the total number of observations
- The cumulative relative frequency adds up the relative frequencies as we move from the first to the last interval

# Plots

- Line plots: plots data against a variable
- Scatter plots: bidimensional plots for bivariate data
- The coordinates of each point are a pair of observations

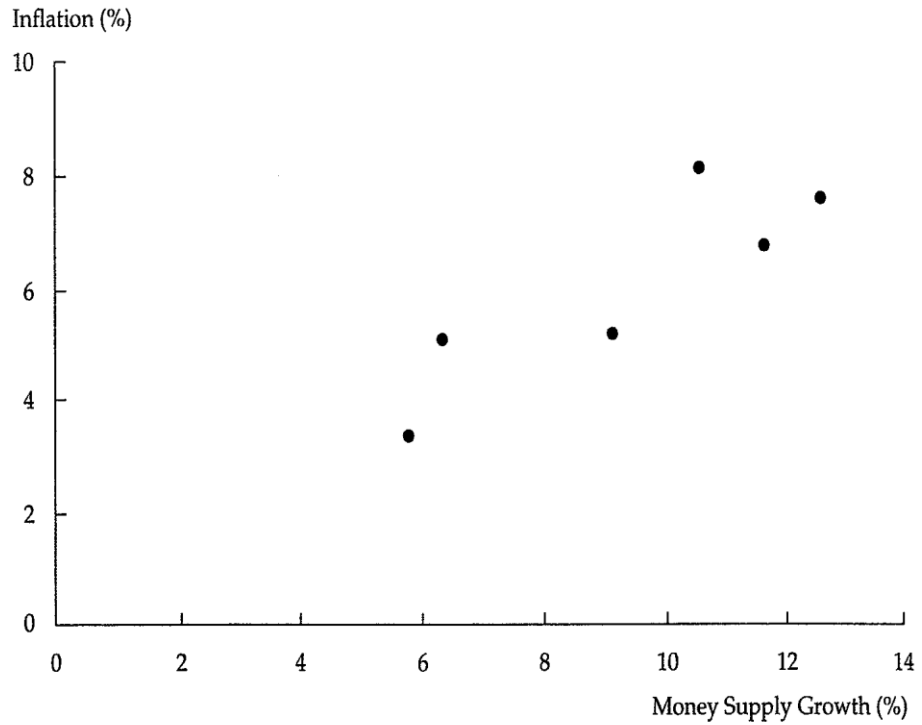
## FIGURE 10-3 Monthly CPI Inflation, Not Seasonally Adjusted




*Source* : Bureau of Labor Statistics.

# Scatterplot of data

**FIGURE 8-1 Scatter Plot of Annual Money Supply Growth Rate and Inflation Rate by Country: 1970–2001**




*Source:* International Monetary Fund

# Where you will find these concepts

- Graphical presentations are used in almost all quantitative financial analysis
- Graphics offer a first intuition for the data
- But above all offer a way to present a synthesis of results
- However, graphics are seldom used as the main tool to draw statistical conclusions

# SECTION 3:
# NUMERICAL DATA ANALYSIS

# Random Variables, samples, and distributions

- In quantitative finance we distinguish between deterministic variables whose value we know with certainty and random variables whose value is uncertain
- We can assume that variables are extracted from some probabilistic models or we can make no assumption about the underlying model
- In both cases we want to characterize our samples with summary indicators
- In this section we introduce parameters that are computed from empirical data
- If we assume a probabilistic model typically indicators correspond to parameters of the underlying probability distribution

# Univariate distribution indicators

- Central Tendency (also called central value)
- Dispersion
- Coefficients of Skewness and Kurtosis
- Percentiles

# Measures of central tendency

- Arithmetic mean

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

- Easy to compute and to manipulate mathematically
- A potential drawback of the mean is its sensitivity to extreme values
- The sum of the deviations around the mean is zero:

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = \sum_{i=1}^{n} X_i - n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = 0$$

# Other measures of central tendency

- Median $\overline{X}_{med}$ : value of the middle item of a sorted set of items

  (odd=(n+1)/2;even=n/2 or (n+2)/2)

- Mode $\overline{X}_{\text{mod}}$: Most frequent value in a distribution

# Measures of location: Quantiles

- Quartiles, Quintiles, Deciles and Percentiles
- The median divides a distribution in half
- Quartiles divide the distribution into quarters, quintiles into fifths, deciles into tenths, percentiles into hundreths.
- The yth percentile is the value at or below which y percent of
- observations lie:

$$L_y = (n+1)\frac{y}{100}$$

- When $L_y$ is not a whole number, use linear interpolation

# Simple measures of dispersion

- Range
- Variance
- Standard deviation

# Measures of dispersion

Definitions

- Range: $RG = \max(X_i) - \min(X_i)$

- (Empirical) Variance: $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

- (Empirical) Standard Deviation: $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

# Importance of measures of dispersion

- In quantitative finance and in asset management measures of dispersion give a summary quantification of risk
- Consider, for example, asset returns
- Variance and standard deviation give an indication of the magnitude of fluctuations of returns around their expected value
- The most common application of standard deviation in finance is the measurement of tracking error with respect to a benchmark
- The management of portfolios of assets is essentially based on finding an optimal trade off between the expectation of returns and their risk quantified by the magnitude of deviations from expectation

# Skewness and Kurtosis

- The sample coefficient of Skewness is a measure of asymmetry of the distribution.

$$S_K = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$$

- The sample coefficient of Excess Kurtosis is a measure of the peakedness of a distribution; gives an idea about the frequency of rare events (fat tails)

$$K_E = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

# Skewness and Kurtosis

For large n Skewness and Kurtosis become:

$$S_K = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{s^3},$$

$$K_E = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{s^4} - 3$$

# Skewness and Kurtosis

- A skewness of + or - 0.5 is unusually large (should be close to zero for a symmetric distribution).
- An excess kurtosis of 1.0 is unusually large (should be close to zero for a normal distribution).

# Importance of skewness and kurtosis

- Variance and standard deviation give an average estimate of the magnitude of fluctuations around the expected value
- Skewness and kurtosis give an additional indication of how fluctuations are distributed
- In fact, different distributions might have the same variance and standard deviation but totally different distribution of fluctuations
- Given any value of standard deviation, fluctuations can be densely packed in some interval around the expected value or they can be spread over much larger intervals (the distribution is then said to have fat tails)
- In addition, fluctuations can be symmetrically distributed around the mean or they might be asymmetrically packed more on one side

# Importance of skewness and kurtosis

- If a distribution has fat tails than very large events occur with a frequency higher than if the distribution were more concentrated around its expected value (See examples of distribution later)
- Tail risk, that is the probability that large events happen, is very important in risk management
- For example, large negative returns, might have serious consequences for financial institutions even if they happen intraday
- Symmetric and asymmetric behaviour is also very important for risk management
- For example, credit risk associated with loans is asymmetric as loans produce relatively small returns with high probability and rare but large losses in case of bankruptcy

# Chebyshev's inequality

- Standard deviation quantifies the dispersion of data
- Therefore we can expect that in any sample a considerable fraction of data are close to the mean
- However we expect that the fraction of data that fall in a given interval around the mean depends on the distribution of data
- A rather surprising result due to Chebyshev states that for any distribution with finite mean and standard deviation we can determine an upper bound to the fraction of data that fall outside any interval centered on the mean and whose width is a given multiple of standard deviation

# Chebyshev's inequality, ctd...

- Chebyshev's Inequality can be stated as follows:
- For any random variable $X$ with finite mean $\mu$ and standard deviation $\sigma$, for any positive real number $k$, the following inequality, called Chebyshev's Inequality, holds:

$$P\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

- That is, the probability that an observation falls more than $k\sigma$ away from the mean is less than $\frac{1}{k^2}$
- For example, the probability than an observation falls more than 2 stds away from the mean is $\frac{1}{2^2} = 0.25$, more than 3 stds is $\frac{1}{2^3} = 0.125$ and so on

# Chebyshev's inequality

- The Chebyshev's Inequality establishes a rather weak upper bound to the probability that an observation falls more than *k* standard deviations away from the mean
- Much smaller upper bounds can be established for peaked distributions
- For example, the probability that a normally distributed variable (to be defined later) falls more than 2 stds away from the mean is 0.05 and more than 3 stds is 0.003

# Multivariate data

- In finance it is particularly important to observe several variables at the same time and try to assess their association
- We will introduce measures of this association such as covariance and correlation
- These two measures are fundamental to understand how to construct diversified portfolios

# Covariance

- The formula for the covariance is:

$$s_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

- The size of the covariance depends on the magnitude of the observations *Xi* and *Yi*
- A large covariance may simply reflect the scale of the variable

# Correlation coefficient

- The covariance depends on the scale of the variables
- The correlation coefficient is a scale-free measure of co-movement obtained by dividing the covariance by the product of standard deviations

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

$$s_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

$$\rho_{XY} = \frac{s_{XY}^2}{s_X s_Y}$$

# Correlation coefficient, ctd...

- Correlation coefficients quantify how two variables move together
- Correlation coefficient is +/-1 if and only if variables are proportional
- In all other cases the correlation coefficient is a number between -1 and +1
- Zero correlation coefficient is an indication that variables do not move together
- However zero correlation coefficient does not imply variables are independent; the latter is true if the variables are normal
- The correlation coefficient is a measure of linear dependence and it does not quantify correctly non linear dependence

# Where you will find these concepts

- Mean, variance, standard deviations and covariance are among the most commonly used statistical concepts in finance, corporate finance, and economics
- Skewness and kurtosis have become important in asset management and risk management
- The Chebyshev's Inequality is a useful bound that can be found in many applications in asset and risk management

# SECTION IV

# COMMONLY USED RANDOM VARIABLES

# Commonly Used Random Variables

- In this section we discuss the properties of a number of commonly used random variables
- By far the most common variable is the normal (also called Gaussian) variable

# Binomial variable

- Consider an experiment that has only two outcomes, conventionally called success or failure or zero and one
- The binomial random variable $X$ represents the number $r$ of successes in $n$ independent trials when the probability of success on each trial is a constant p.
- We write $X \sim B(n, p)$
- The variable $X$ can assume n+1 possible values: 0,1,2, ⋯,n.

# Binomial distribution

- The probability of r successes is:

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, \quad 0 \le r \le n$$

- The following properties hold:

$$E(X) = np$$

$$\text{var}(X) = np(1-p)$$

$$\sigma_X = \sqrt{np(1-p)}$$

# Poisson distribution

- The Poisson distribution is defined as:

$$P(X = r) = \frac{(\lambda t)^r e^{-\lambda t}}{r!}$$

- With the following properties:

$$E(X) = \mu = \lambda t$$

$$\text{var}(X) = \mu$$

- The Poisson distribution describes the number of independent events that happen in a given interval *t* (Poisson process)
- It is a first approximation for describing the number of trades that happen in a given interval in stock Exchanges; also used to approximately describe the arrival of insurance claims

# Uniform continuous distribution

Constant density in an interval a,b:

$$f(x) = \frac{1}{b-a},\ a \le x \le b$$

$$0,\ x < a, x > b$$

$$U \sim Uniform(a,b)$$

$$E(U) = \frac{a+b}{2}$$

$$\mathrm{var}(U) = \frac{(b-a)^2}{12}$$

# Application

- The uniform distribution is used extensively in simulation applications.
- A random sample is drawn from a uniform distribution on the interval [0,1].
- The inverse of the cdf is applied to this random sample in order to obtain a random sample from a continuous distribution.

# Normal distribution

- The normal distribution is the most widely used probability distribution in finance.
- It is due to the remarkable result known as Central Limit Theorem
- It says that, for any variable with finite variance, the mean of a large number of independent realizations of that variable is approximately normally distributed.
- The distribution is completely characterized by its mean and
- standard deviation.
- The density function is symmetrical and tails are thin.
- It is in general a good approximation for monthly returns on an
- equity index.

# Normal distribution

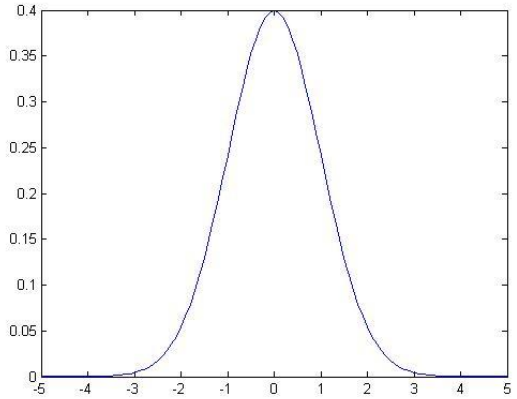
- Density function: $X \sim N(\sigma, \mu)$

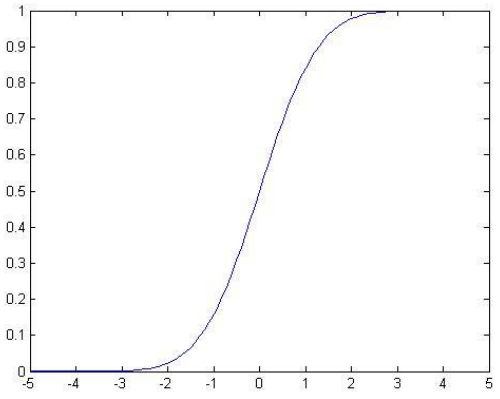
$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

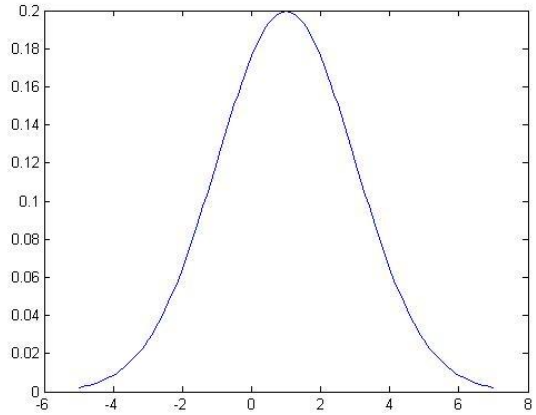
- Standard normal variable Z has mean zero and std=1
- We standardize by defining:

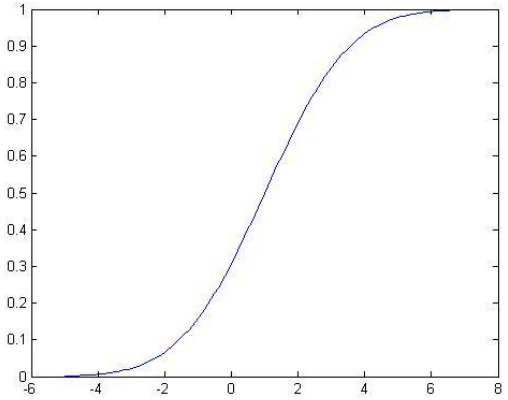
$$Z = \frac{X - \mu}{\sigma}$$

# Normal distribution: mu =0, sigma =1 and mu=1, sigma =2



# Lognormal distribution

- A variable is said to be lognormally distributed if the natural logarithm of the variable is normally distributed
- Density:

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2}}{\sigma x\sqrt{2\pi}}$$

- where μ is the mean and $\sigma$ the standard deviation of the underlying normal distribution
- If we assume that monthly logreturns are approximately normal and not serially correlated then prices have a lognormal distribution

# Lognormal distribution, ctd...

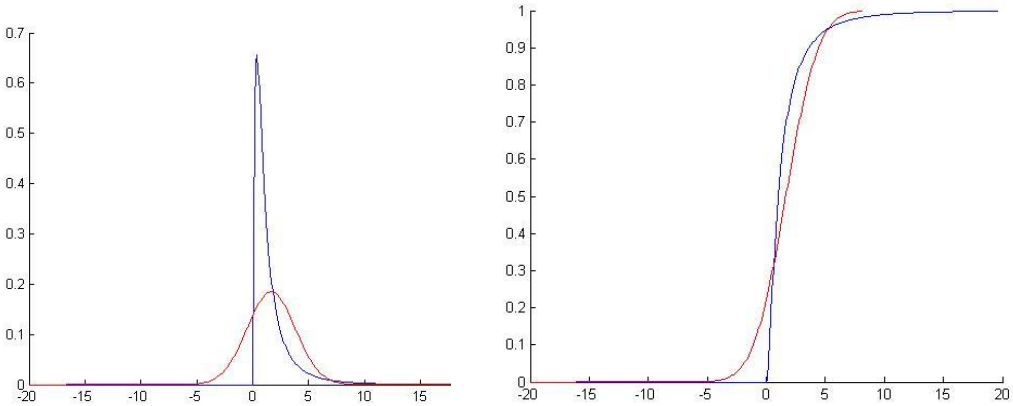
- The mean and variance of the lognormal distribution are given by the following expressions:

$$m = E(X) = e^{\mu+\frac{1}{2}\sigma^2}$$

$$v = \text{var}(X) = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}$$

- The lognormal distribution is zero for negative values of the variable $x$, that is,
- A random variable with a lognormal distribution assumes only positive values (example stock prices)

Lognormal distribution mu=0, sigma=1 blue, normal distribution same mean and std red



# Chi-square

- The square of a standard normal variable is distributed as a $\chi^2$ with one degree of freedom.

$$Z^2 = \chi^2_{(1)}$$

- If $Z_1, Z_2, \ldots, Z_k$ are k independent standard normal variables, then their sum is said to be distributed as a chi-square with *k* degrees of freedom

$$\sum_i Z_i^2 = \chi^2_k$$

# Density and degrees of freedom of the chi-square variable

- Density:

$$p(x,k)=\begin{cases} Cx^{\frac{k}{2}-1}e^{-\frac{x}{2}},\ C=\dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} & \text{for } x\geq 0 \\ 0 \text{ otherwise} \end{cases}$$

- The term degree of freedom refers to the number of independent observations in a sum of squares.
- In general it is the number of free observations to compute a quantity.
- For example, in the formula to compute the sample variance, we lose one degree of freedom to compute the mean of X .

# Properties of chi square

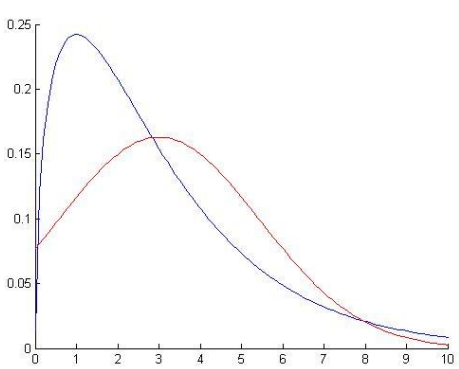
- It is represented only in the positive values since variables are squares.
- The distribution is asymmetric. The degree of asymmetry depends on the number of degrees of freedom.

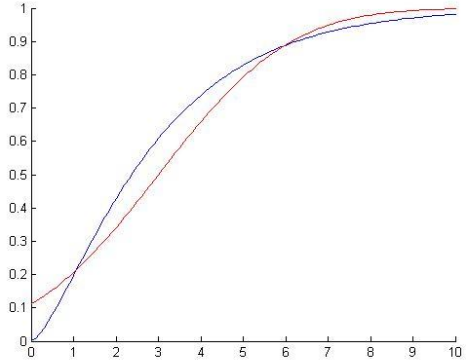
  $E(\chi^2) = k$

  $\text{var}(\chi^2) = k$

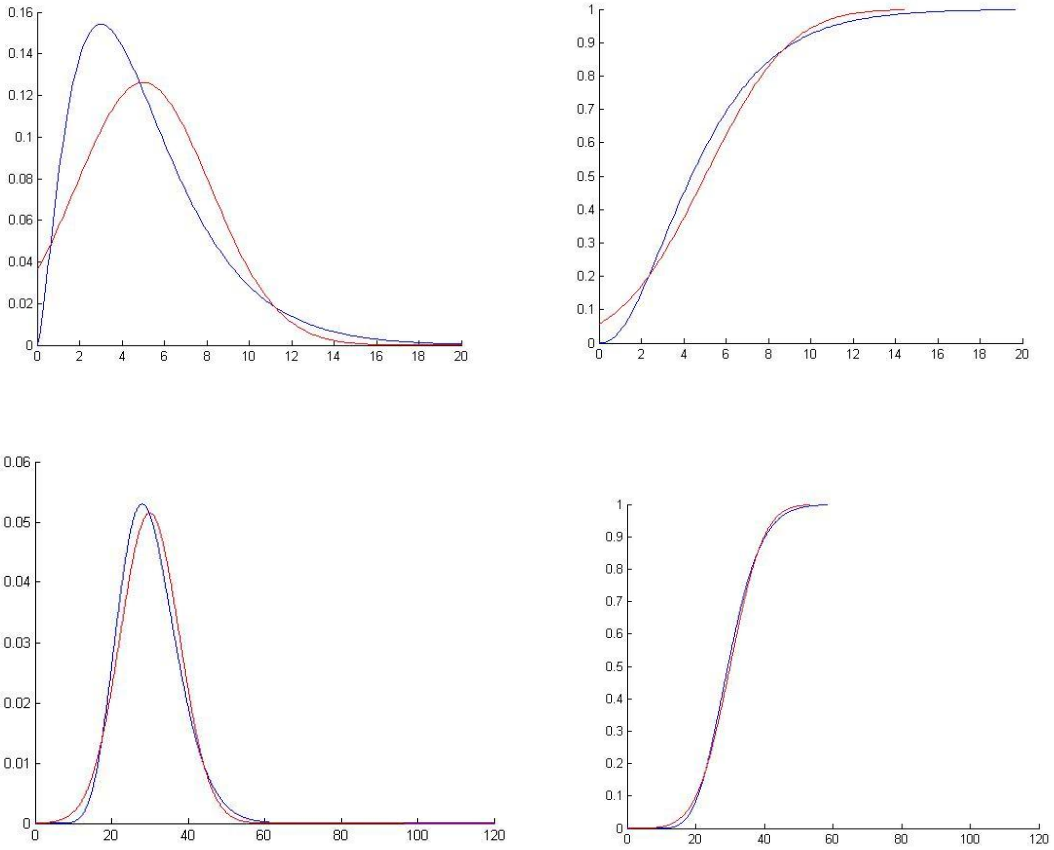
- If $Z_1, Z_2$ are 2 chi-square variables with $k_1$ and $k_2$ degrees of freedom, the variable $Z_1 + Z_2$ is distributed as a chi-square with $k_1 + k_2$ degrees of freedom
- The chi-square distribution is used in many statistical tests and in the analysis of variance

# Chisquare distribution 3df blue, normal distribution same mean and std red



# Chisquare distribution 5 and 30 df blue, normal distribution same meand and std red



# Student's t Distribution

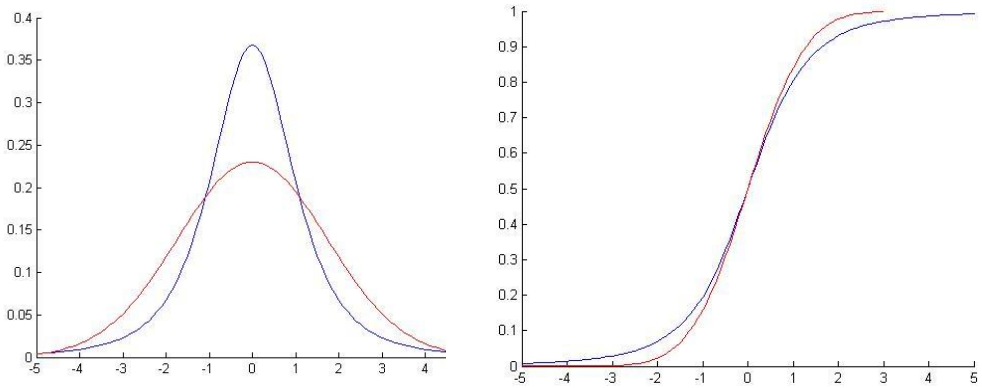
- The Student's t distribution is used very often as a sampling distribution or to characterize non-normality
- It is the distribution of the ratio $Z\sqrt{\frac{\nu}{V}}$ where $Z$ is a standard normal variable and $V$ is a chi-square variable with $\nu$ degrees of freedom and $Z$ and $V$ are independent
- Density:

$$p(x) = C\left(1+\frac{x^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}, \; C = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}$$

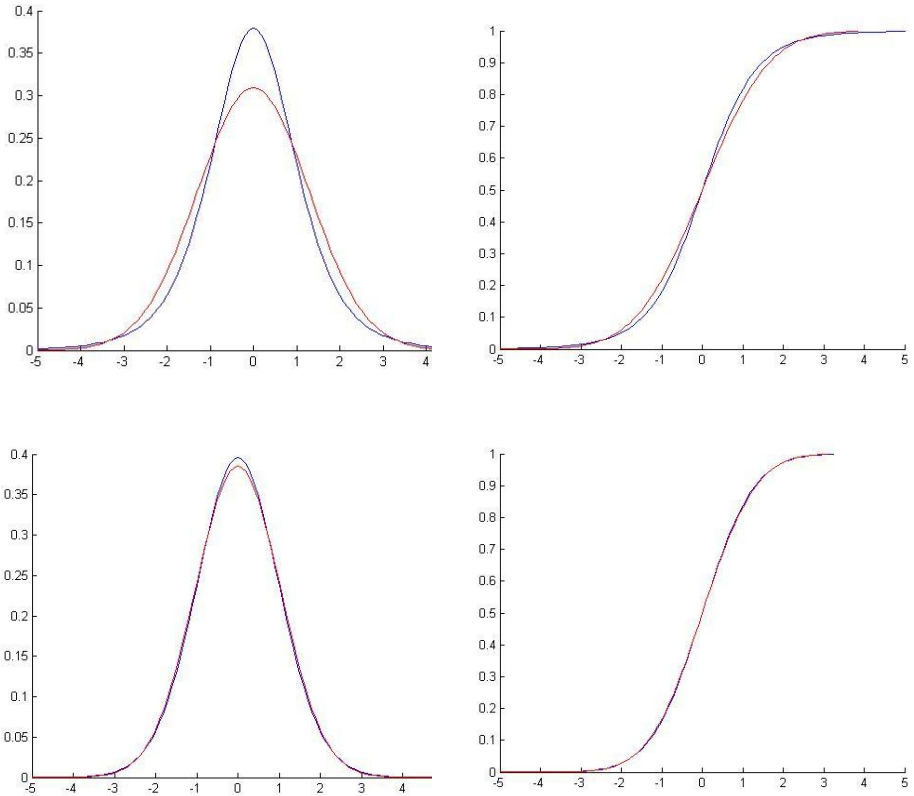
# Properties

- As the normal distribution, the *t* distribution is symmetric.
- The mean is zero, but the variance is $k/(k-2)$ (the variance of the distribution is defined for $k > 2$, where k is the number of degrees of freedom).
- As k becomes large, the distribution tends towards a normal distribution.
- The variance is then very close to 1.

# Student's t distribution 3df blue, normal distribution same mean and std red



# Student's t distribution 5df and 30df blue, normal distribution same mean and std red



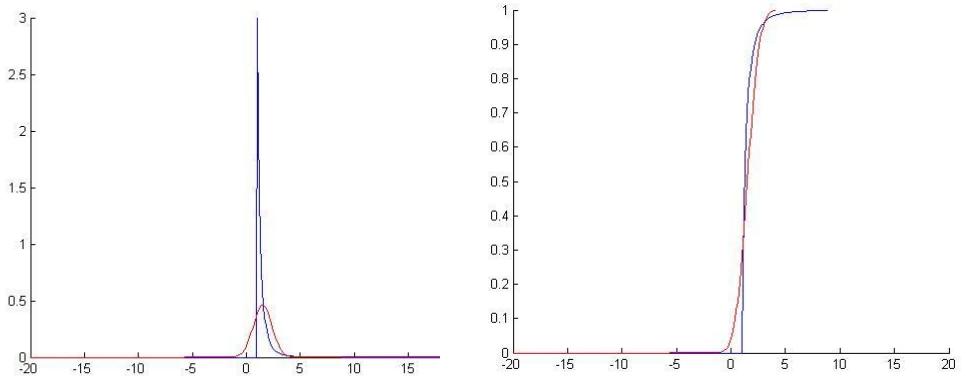
# Pareto distribution

- The Pareto distribution is a fat-tailed distribution
- A variable $X$ has a Pareto distribution if it has the following density:

$$p(x)=\begin{cases} x^{-(\alpha+1)}, & \mathrm{x}\geq 1 \\ 0, & \mathrm{x}<1 \end{cases}$$

- Where $\alpha$ is a positive tail parameter that determines the weight of the tails
- If $\alpha \leq 1$ the expectation is infinite, if $\alpha>1,\ E(\mathrm{X})=\dfrac{\alpha}{\alpha-1}$
- If $\alpha \leq 2$ the variance is infinite, if $\alpha>2,\ \mathrm{var}(\mathrm{X})=\left(\dfrac{1}{\alpha-1}\right)^2\dfrac{\alpha}{\alpha-2}$

# Pareto distribution alpha = 3 blue, normal distribution same mean and std red



# Tail risk

- Tail risk is defined as the probability that the returns of an asset or a portfolio have fluctuations larger than 3 standard deviations
- Tail risk is often stated relative to a normal distribution, in the sense that there is tail risk if the probability of movements in excess of 3 standard deviations exceeds that of a normal distribution
- To give some intuitive meaning to the notion of tail risk in the following table we compare the probability that a variable $X$ takes values larger than 1,2,3,4,5,6 standard deviations to the right of the mean for different distributions

# Tail risk, ctd...

- We compare normal, lognormal, chi-square, Student's t and Pareto distributions
- Note that the tail risk of the normal distribution is independent from the mean and the variance of the distribution
- This is not true for the lognormal distribution
- The chi-square and *t* distributions are one parameter distributions where the mean and the variance depend on the number of degrees of freedom
- The Pareto distribution is a special case of a more general family of distributions where the mean and variance depend on several parameters

| | P(X>μ+1σ) | P(X>μ+2σ) | P(X>μ+3σ) | P(X>μ+4σ) | P(X>μ+5σ) | P(X>μ+6σ) |
|---|---|---|---|---|---|---|
| Normal | 0.1587 | 0.0228 | 0.0013 | 0.000031 | 0.00000029 | 0.0000000001 |
| Lognormal | 0.0905 | 0.0370 | 0.0180 | 0.0099 | 0.0058 | 0.0037 |
| Chi-square 3df | 0.1417 | 0.0481 | 0.0158 | 0.0051 | 0.0016 | 0.0005 |
| Chi-square 5df | 0.1475 | 0.0453 | 0.0128 | 0.0034 | 0.0009 | 0.00022 |
| Student's t 3df | 0.0908 | 0.0203 | 0.0069 | 0.0031 | 0.0016 | 0.00095 |
| Student's t 5df | 0.1256 | 0.0247 | 0.0059 | 0.0018 | 0.0007 | 0.00028 |
| Pareto alpha = 3 | 0.0755 | 0.0296 | 0.0145 | 0.0082 | 0.0050 | 0.0033 |

# SECTION V:
# SAMPLING AND ESTIMATION

# Sampling and Estimation

- Sampling is the process of obtaining a sample. We want to obtain information on a population (value of a parameter) through samples (estimation of a parameter by using sample statistics).
- Two key elements
- The central limit theorem: we can make statements about the population mean, even when the distribution of a random variable is unknown, through the limit distribution of the sample mean.
- Estimation: methods for using a sample to estimate a parameter.

# Sampling

- Simple random sampling
- A simple random sample is a subset of a larger population created in such a way that each element of the population has an equal probability of being selected in the subset
- Sampling error is the difference between the observed value of a statistic and the quantity it is intended to estimate

# Sampling distribution of a statistic

- Sample statistics, such as the sample mean, computed on the basis of a random sample, are valid estimates of the underlying population parameters.
- A sample statistic is a random variable. It has a distribution.

- The sampling distribution of a statistic is the distribution of all the distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population

# Distribution of the sample mean

- The sample mean is a function of the random outcomes of a random variable
- The sample mean is itself a random variable with a probability distribution.
- That probability distribution is called the statistic's sampling distribution. It tells how closely we can expect the sample mean to match the population mean.
- The central limit theorem tells us what the sampling distribution is in many cases of interest

# The central limit theorem

- Given a population described by any probability distribution with mean $\mu$ and finite variance $\sigma^2$
- The sampling distribution of the sample mean X computed from samples of size n from this population
- will be approximately normal with mean $\mu$ and variance $\frac{\sigma^2}{n}$ when the sample size n is large
- NOTE: the central limit theorem does not hold if the variance of the distribution is infinite

# Point estimation

- Suppose we have a sample from a population
- We want to estimate some population parameters from the sample
- We need to compute a point estimator, that is, a number which is close to the true parameter to be estimated
- A point estimator is a function of the sample data
- An estimator is a random variable

# Point Estimation

Example

- Assume we know that a population is distributed normally but we do not know its mean and variance
- We draw a sample and we use the following rules to estimate the mean and variance

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

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

- These two formulas are called estimators of the mean and variance of the
- population.
- Once we replace the $Xi$ by the values drawn in the sample, we obtain estimates
- of the mean and variance of the population.
- We say we have carried out a point estimation of the parameters.
- We can have several point estimators of the parameters. We will need criteria to choose among them

# Properties of point estimators

- When we were looking for an estimator of the mean $\mu$ of a distribution, we proposed the sample mean X. Why not the median or the mode, that we saw were other measures of the central tendency of a distribution.
- In practice, the sample mean has several properties thought to be desirable in statistics: linearity, unbiasedness, and efficiency.
- We will define these properties and state that the sample mean is the best linear unbiased estimator of the parameter $\mu x$.

# Properties of Point Estimators

Linearity

- An estimator is linear if it is a linear function of observations.

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

- A linear estimator is easier to process than a nonlinear estimator

# Properties of point estimators

Unbiasedness

- An estimator is said to be unbiased if, in average, it coincides with the true value of the parameter.

  $E(X) = \mu$

- Otherwise, the estimator is said to be biased. To determine if an estimator is biased, we need therefore to compute its expectation and verify if it is equal to the true value of the parameter
- Unbiasedness is a theoretical property of the estimators
- An unbiased estimator is preferred since in average it will give the true value of the parameter.

# Properties of Point Estimators

Efficiency

- How to compare several unbiased estimators?
- The sample mean estimator

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

- The sample median estimator

$$X_{med} \sim N\left(\mu, \frac{\pi}{2}\frac{\sigma^2}{n}\right)$$

- These are two unbiased estimators of the population mean μ. But we notice that the median has a larger variance than the sample mean.

$$\frac{\text{var}(X_{med})}{\text{var}(\bar{X})} = \frac{\pi}{2} = 1.571$$

- We will prefer $\bar{X}$ to $X_{med}$ because of its lower variance, since it will estimate the parameter with more precision (the confidence interval built for μ will be smaller).
- We will say that X is an efficient estimator (or best unbiased estimator).

# Consistency

- Unbiasedness and efficiency are properties of an estimator's sampling distribution that hold for any size sample.
- In some problems, we cannot find estimators that have such desirable properties as unbiasedness in small samples.
- One solution is to look for estimators that have good properties in extremely large sample, so-called asymptotic properties.
- The most important one is consistency

# Definition of consistency

- A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.
- A consistent estimator is an estimator whose sampling distribution becomes concentrated on the value of the parameter it is intended to estimate as the sample size approaches infinity.
- The sample mean is both an efficient estimator and a consistent estimator of the population mean: as the sample size n goes to infinity, its standard error s pn goes to zero and its sampling distribution gets concentrated right over the value of the population mean μ.

# Best Linear Unbiased Estimator

- In econometrics we will often refer to this property
- It will meet the three criteria we just saw
- It will be the best (smallest variance) estimator among the unbiased and linear estimators

# Confidence Intervals

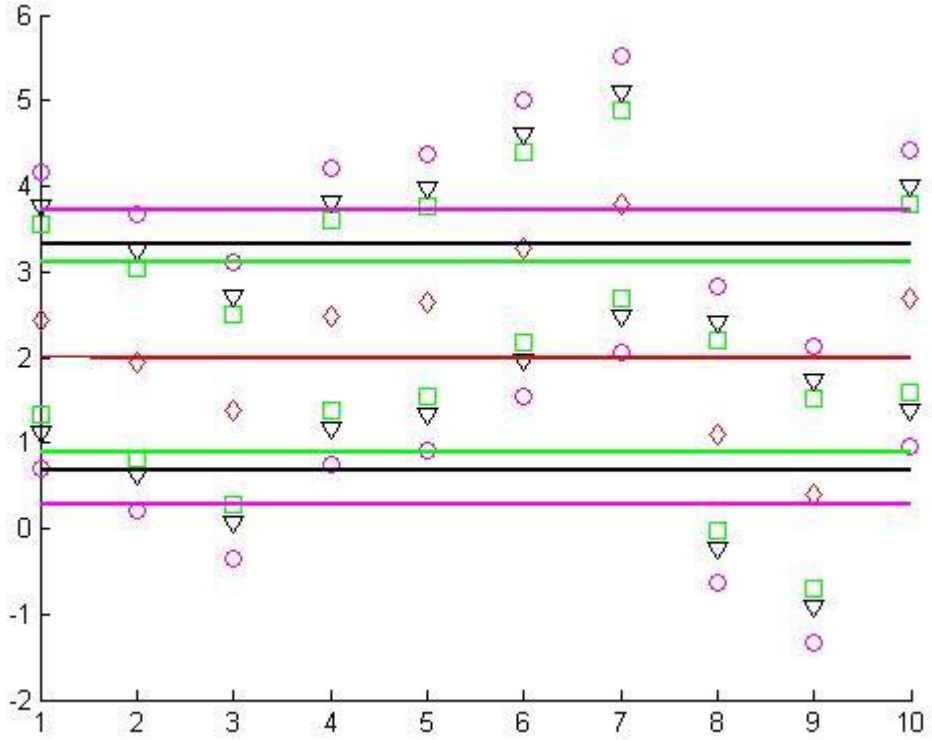
- How can we trust a single estimate of a parameter?
- For example, how can we trust a single estimate of the mean parameter?
- Will not it be better to say that the interval 9 to 17 contains most likely the true mean of the population instead of saying that 13 is the best estimated value of the mean of the population of price-dividend ratios
- This is the idea of interval estimation based on the concept of sample distribution

# Confidence Intervals

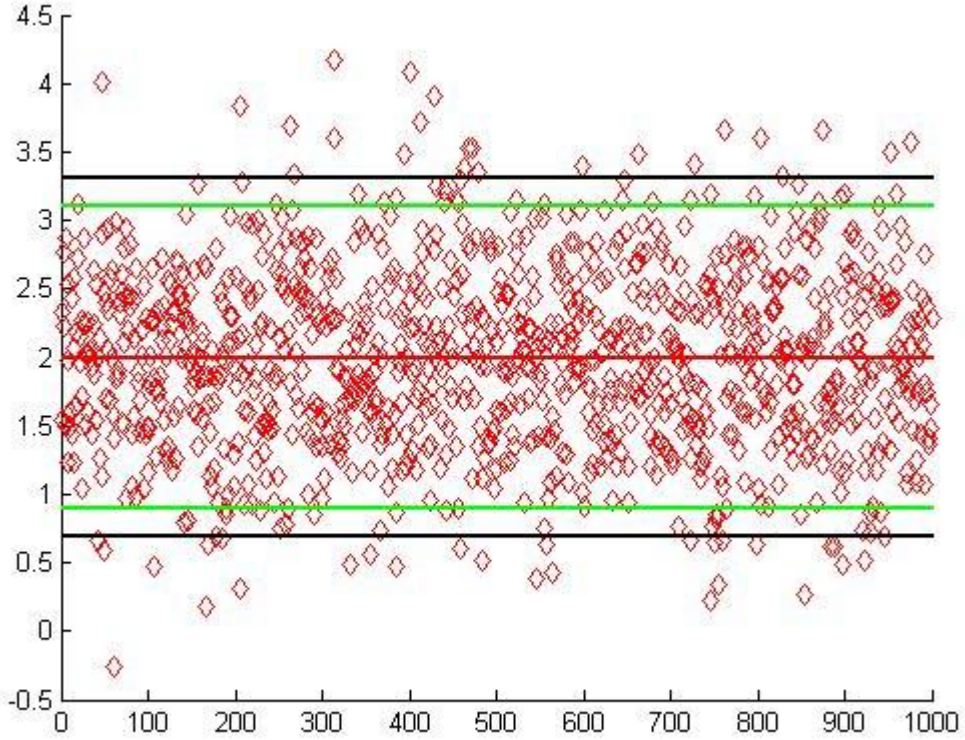
Definition

- A confidence interval is a range for which one can assert with a given probability $(1-\alpha)$ , called the degree of confidence, that it will contain the parameter it is intended to estimate.
- This interval is often referred to as the $(1-\alpha)\%$ confidence interval of the parameter
- NOTE: A confidence interval is a random interval around an estimated parameters; it changes with each estimate of the parameter
- These intervals should be interpreted as follows: if such intervals are built a large number of times, by drawing repeated samples, $(1-\alpha)\%$ of the intervals will include the true mean

# Confidence intervals 5,10
# Percentages: 0.7000, 0.8000, 0.9000



# Distribution 5,1000
# Percentages: 0.9020, 0.9500, 0.9920



# SECTION VI

# LINEAR REGRESSION

# Linear Regression with One Independent Variable

- Linear regression  quantifies the strength of the linear relationship between two variables…
- Tests hypotheses about the relation between two variables…
- Uses one variable to predict another variable.
- It assumes a linear relationship between the dependent and the independent variable

$$Y = b_0 + b_1 X + \varepsilon$$

- Y is the dependent variable and X the independent variable.
- b0, the intercept, and b1, the slope coefficient, are called the regression coefficients.
- The error term represents the portion of the dependent variable that cannot be explained by the independent variable.
- Cross-sectional and time series regressions.

# Regression model definition

- Simple regression: $$Y = b_0 + b_1 X + \varepsilon$$

- Multiple regression: $$Y = b_0 + b_1 X_1 + \cdots + b_K X_K + \varepsilon$$

- Design matrix: $$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & \cdots & X_{k,n} \end{bmatrix}, b = \begin{bmatrix} b_0 \\ \vdots \\ b_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- Regression in matrix form: $$Y = Xb + \varepsilon$$

# Interpretation of the Error Terms

- For a same $X_i$ we can have a different $Y_i$ . The relationship is an approximate one.
- Even though we could use a lot of variables (not recommended based on a principle of model parsimony) a model is always an approximation.
- Sometimes (survey data) the error term can capture some measurement error.
- It could also be an approximation error introduced by the choice of a linear form while the true relationship is nonlinear.
- Contrary to theYi these errors are unobservable. Once we estimate the parameters we will have estimated values for these residual errors.

# Estimation of the linear regression model

The estimation computes a line that best fits the observations by choosing values for the b0 and b1 coefficients in order to minimize the sum of squared vertical distances between the observations and the regression line

# Estimation simple regression (one independent variable)

- Method of the Least Squares (LS)

- Given a sample $Y_i, X_i, i = 1,2,\ldots,n$

- Minimize the sum of the squares of the differences $(Y_i - b_0 - b_1 X_i)^2$

$$\min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \min_{b_0,b_1} \sum_{i=1}^{n} \varepsilon_i^{\,2}$$

# Estimators of the regression coefficients

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ sample means
- $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2, s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ sample variances
- $s_{XY} = \mathrm{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$ sample covariance
- $\hat{b}_0 = \bar{Y} - \bar{X}\hat{b}_1$ estimators of the regression coefficients

$$\hat{b}_1 = \frac{s_{XY}}{s_X^2}$$

# Estimation multiple regression

- Least square estimators: minimize the sum of squared errors

- Applying First Order Conditions (partial derivatives equal to 0) we obtain

$$\min_b\left(\varepsilon'\varepsilon\right)=\min_b\left(\left(Y-Xb\right)'\left(Y-Xb\right)\right)$$

$$\frac{\partial\left(\left(Y-Xb\right)'\left(Y-Xb\right)\right)}{\partial b}=X'\left(Y-Xb\right)$$

$$X'Y=X'Xb$$

$$\hat{b}=\left(X'X\right)^{-1}X'Y$$

# If only one variable:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & n\overline{X} \\ n\overline{X} & \sum_{i=1}^{n} X_i^2 \end{bmatrix},$$

$$\begin{bmatrix} n & n\overline{X} \\ n\overline{X} & \sum_{i=1}^{n} X_i^2 \end{bmatrix}^{-1} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - n^2\overline{X}^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 & -n\overline{X} \\ -n\overline{X} & n \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \overline{Y} \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - n^2\overline{X}^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 & -n\overline{X} \\ -n\overline{X} & n \end{bmatrix} \begin{bmatrix} \overline{Y} \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - n^2\overline{X}^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 \overline{Y} - n\overline{X}\sum_{i=1}^{n} X_i Y_i \\ -n^2\overline{XY} + n\sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

# Remarks

- The regression line goes through the center of gravity
- The slope coefficient is slightly different from the coefficient of correlation and is not bounded

$$\hat{b}_1 = \rho \frac{s_Y}{s_X}$$

- The estimators b0 and b1 are random variables.

# Where do we find linear regressions

- Linear regressions are arguably the most common model in finance
- We find regression in asset management, in corporate finance, in economics
- Every time we argue that two variables are linearly dependent
- For example, performance measurement in asset management makes large use of regressions over benchmarks