Portfolio Theory

Dr. Andrea Rigamonti

andrea.rigamonti@econ.muni.cz

Lecture 3

Content:

- Statistics notions

Statistics notions

A **random variable** (r.v.) is one whose value depends on the outcome of a random experiment.

- A **discrete** r.v. can only take certain specific values.
- A **continuous** r.v. can take any value (inside a certain interval)

A **probability distribution** is a function that assigns a certain probability that each possible outcome will occur.

- For a discrete r.v. a **probability mass function** (or probability distribution function) is the function that links each outcome with the corresponding probability.
- For a continuous r.v. we have a **probability density function** (pdf), and the probability of the events is given by the area under a function (i.e. by its integral).

Statistics notions

The most popular continuous distribution is the **Gaussian** (or **normal**) distribution. Characteristics:

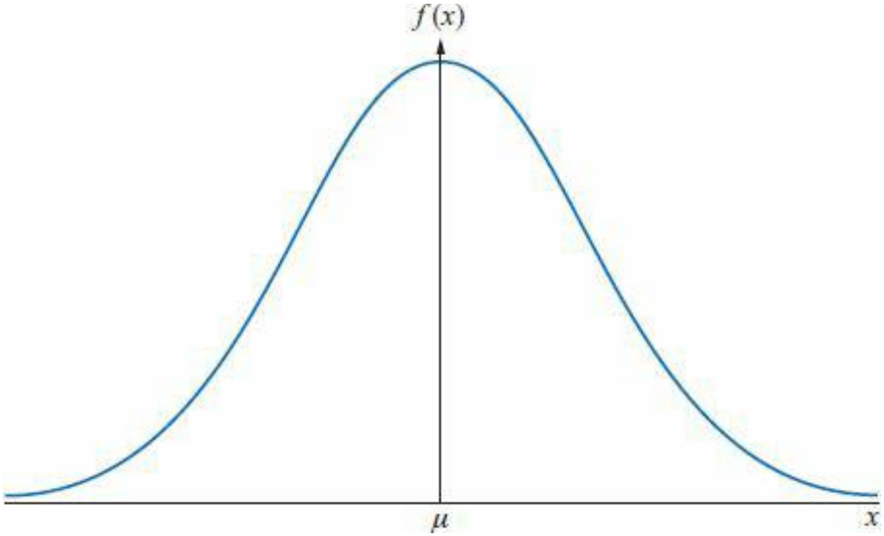
- Symmetric
- Unimodal
- Completely described by its mean μ and variance σ^2
- Any linear transformation of a normally distributed r.v., or any linear combination of independent normally distributed r.v., is itself normally distributed.

The pdf of a normally distributed r.v. is given by:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

Statistics notions

The pdf of a normal r.v. has a bell shape:



- The area under the pdf measures the probability
- Probability is measured in the range [0,1], so the entire area under the curve (the sum of probability of all events) is 1
- The probability of an exact value is zero (a line has area 0)
- We can compute the probability of an interval

Statistics notions

A **standard normal** distribution is a normal distribution with mean μ $= 0$ and standard deviation $\sigma = 1$.

Given a normally distributed r.v. Y, a standard normally distributed r.v. is given by:

$$Z = \frac{y-\mu}{\sigma} \sim N(0,1)$$

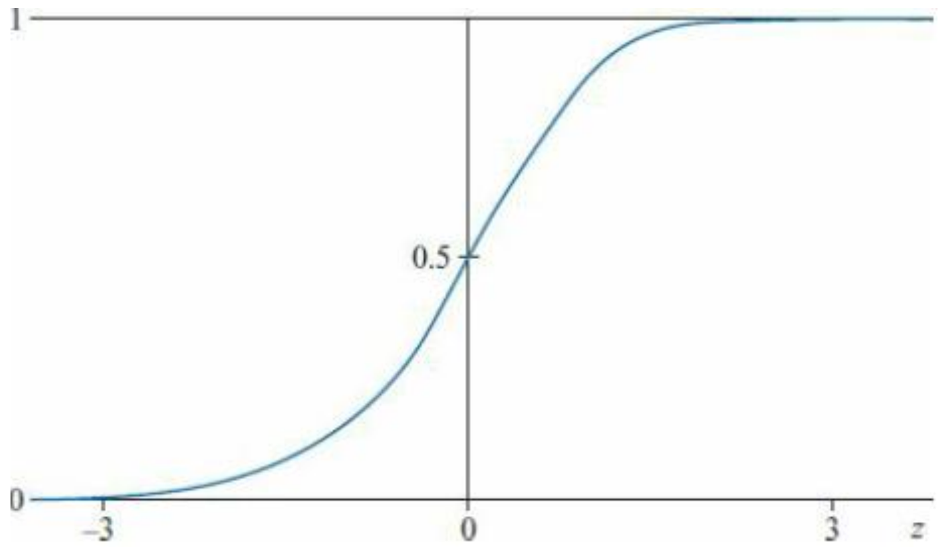
The **cumulative distribution function** (or cumulative density function), cdf, of a continuous r.v. measures the probability that the r.v. is less (or equal to) a certain value.

The cdf $F(y)$ and is given by the integral of the pdf:

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)\, dt$$

Statistics notions

The cdf of a normally distributed r.v. has a sigmoid shape:



- It is given by:

$$\mathrm{F}(y) = P(Y \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}\, dy$$

- There are no exact analytical solutions for this integral
- The values of the cdf is therefore computed numerically for the standard normal and provided into tables

Statistics notions

The average value of *n* financial returns R_i is known as measure of location or **measure of central tendency**.

- Mode: the most frequently occurring value
- Median: the middle value when the elements are arranged in ascending order
- Arithmetic mean:

$$\overline{R_A} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

- Geometric mean:

$$\overline{R_G} = \sqrt[n]{\prod_{i=1}^{n}(1 + R_i)} - 1$$

Statistics notions

With log returns $\overline{R_A}$ and $\overline{R_G}$ are the same

With simple returns $\overline{R_G} < \overline{R_A}$

The geometric mean accounts for compounding and is less affected by outliers BUT it <u>cannot</u> be used as estimate for future returns

The most important **measures of spread** are:

- The range: *max – min*
- The interquartile range: $Q_3 - Q_1$
  where Q_3 is the $\frac{3}{4}(n+1)^{th}$ value and Q_1 is the $\frac{1}{4}(n+1)^{th}$

Statistics notions

- The variance:

$$\overline{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \overline{R_A})^2$$

- The standard deviation σ (square root of the variance)
- The coefficient of variation: $CV = \frac{\sigma}{\mu_A}$

The bar in $\overline{\sigma^2}$ indicates a <u>(sample) estimate</u>. We divide by $n-1$ to correct for the loss of a degree of freedom. The formula for the <u>population</u> variance is:

$$\sigma^2 \ = \frac{1}{n}\sum_{i=1}^{n}(R_i - \mu_A)^2$$

where μ_A is the “true” mean.

Statistics notions

In practice, true parameter values are generally not available and we use estimates. To keep the notation light, the bar is usually not used when there is no risk of confusion.

Higher moments are important when data are not gaussian

- The skewness measures the asymmetry around the mean:

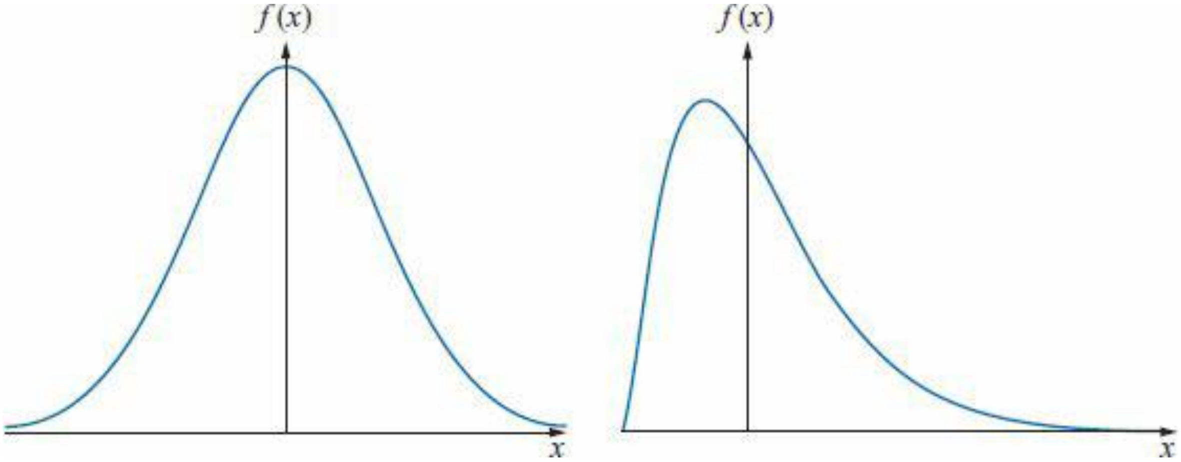
$$s = \frac{1}{\sigma^3}\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \overline{R_A})^3$$

- The kurtosis measures the fatness of the tails and how peaked at the mean the distribution is:

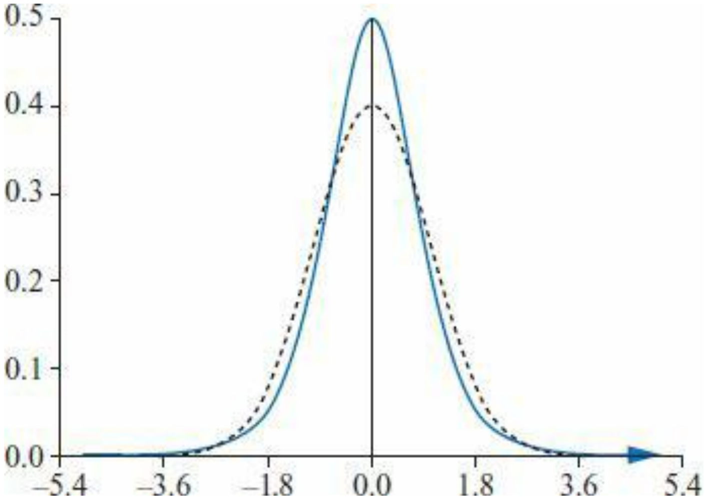
$$k = \frac{1}{\sigma^4}\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})^4$$

Statistics notions

Normal vs positive (or right) skewed distribution



Normal (k=3) vs leptokurtic (k>3) distribution



Statistics notions

Measures of association evaluate links between variables:

- The covariance between two variables X and Y is:

$$\sigma_{X,Y} = Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

- The (Pearson) correlation between X and Y is:

$$\rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

A negative (positive) covariance means that the variables move, on average, in the opposite (same) direction.

Statistics notions

The correlation takes value between -1 (perfect negative correlation) and 1 (perfect positive correlation).

$\sigma_{X,Y}$ or $\rho_{X,Y}$ equal to zero indicate no linear correlation, but not necessarily independence.

Only if X and Y are joint normally distributed, zero covariance implies that they are independent.

Statistics notions

The variance-covariance matrix, or simply **covariance matrix**, is a symmetric matrix that contains all the variances (on the diagonal) and covariances (off the diagonal) of the data on which it is estimated:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1n} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \cdots & \sigma_n^2 \end{bmatrix}$$

(Σ is the standard notation to indicate the covariance matrix; do not confuse it with the sum symbol!)

Statistics notions

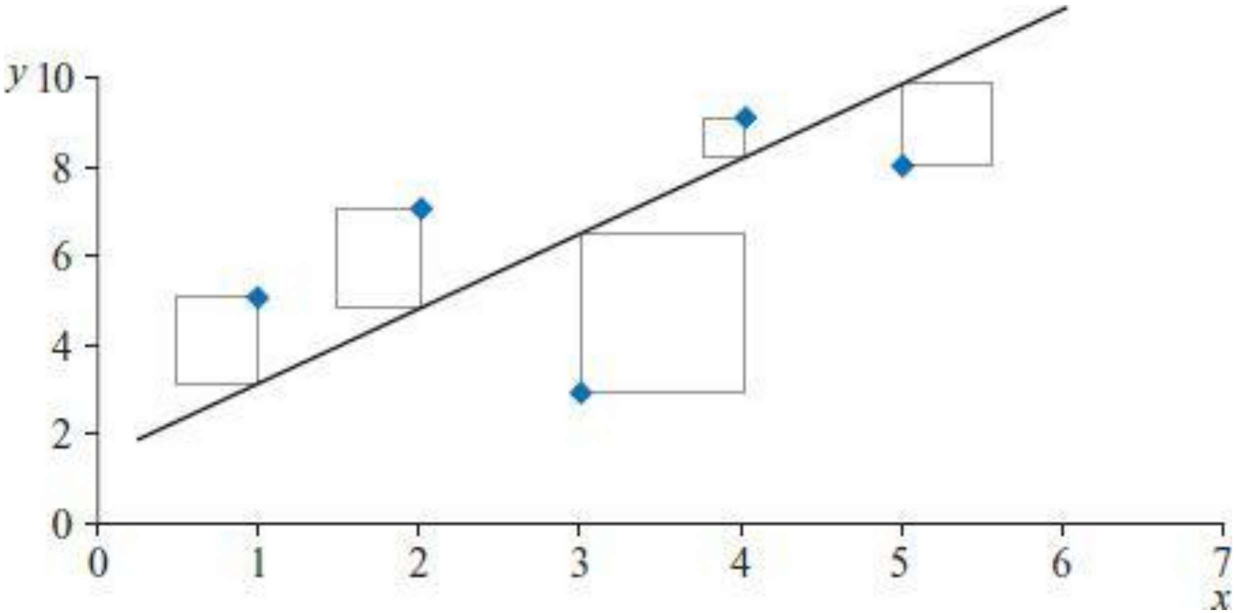
The relationship between a dependent variable and an explanatory variable can be described with the **linear regression model with a single regressor:**

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

This equation describes a line that fits the data, plus a disturbance (or error) term ε_i.

To fit the data we choose the parameters α and β using the **Ordinary Least Squares (OLS)**: take each vertical distance from the point to the line, square it and then minimize the total sum of the areas of squares.

Statistics notions



- For each data point i we denote the fitted value obtained from the regression line as $\hat{y}_i$.
- $\hat{\varepsilon}_i$ denotes the residual: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
- The OLS method minimizes the residual sum of squares (RSS):

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Statistics notions

The values of α and β selected by minimizing the RSS are $\hat{\alpha}$ and $\hat{\beta}$.

The equation of the fitted line is therefore:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

The OLS estimators for the single regressor case are:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

If $E(\varepsilon_i) = 0, var(\varepsilon_i) = \sigma^2 < \infty, cov(\varepsilon_i, \varepsilon_j) = 0, cov(\varepsilon_i, x_i) = 0,$ then estimators $\hat{\alpha}$ and $\hat{\beta}$ determined by OLS are BLUE: Best Linear Unbiased Estimator.

Statistics notions

The **multiple linear regression model** with k regressors (and $k-1$ explanatory variables) has the form

$$y_t = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

- Each $\beta_1, \beta_2, \ldots, \beta_k$ is known as **partial regression coefficient**, and represents the partial effect of the given explanatory variable on the explained variable, after holding constant (or eliminating the effect of) all other explanatory variables.
- The regressor x_{1i} is not written explicitly, as it is always equal to 1.
- The intercept $\beta_1 x_{1i} = \beta_1$is not an explanatory variable, although for notational convenience k might be referred to as the "number of explanatory variables".

Statistics notions

The model can be written in matrix form as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are a n x 1 vectors, $\boldsymbol{\beta}$ is a k x 1 vector, and $\boldsymbol{X}$ is a n x k matrix.

- Written in extended form, the equation describing the model is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- The RSS has to be minimized with respect to all the elements of $\boldsymbol{\beta}$

$$RSS = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = [\hat{\varepsilon}_1 \quad \hat{\varepsilon}_2 \quad \cdots \quad \hat{\varepsilon}_n] \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{bmatrix} = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \cdots + \hat{\varepsilon}_n^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Statistics notions

- The coefficient estimates are given by

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{y}$$

- The variance of the error terms is given by $s^2 = \frac{\hat{\boldsymbol{\varepsilon}}\prime\hat{\boldsymbol{\varepsilon}}}{n-k}$. We have $n-k$ degrees of freedom because k parameters were estimated.

- The covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$var\left(\widehat{\boldsymbol{\beta}}\right) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- The standard errors of the coefficients in $\widehat{\boldsymbol{\beta}}$ are thus given by the square roots of the terms on the leading diagonal of $var\left(\widehat{\boldsymbol{\beta}}\right)$.