Measures of Data Distributions

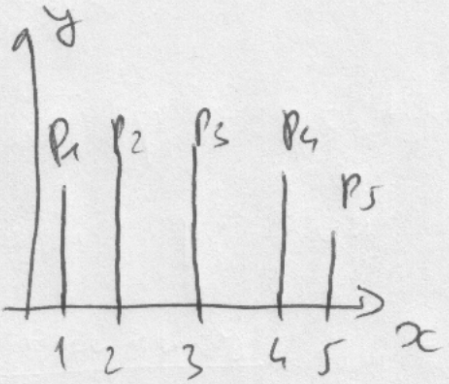
- What is random variable?
- Suppose x is a random quantity: it can take on any one of a finite set of numbers, say $\{x_1, x_2, .., x_m\}$; Assume further that associated with each possible x_i, there is a probability p_i that represents the relative chance of an occurance of x_i
- The p_i's satisfy $\sum_{i=1}^{m} p_i = 1$, and $p_i \geq 0$
- Each p_i can be thought of as the <u>relative frequency</u> with which x_i would occur if an experiment of observing x were repeated infinitely often
- The variable x characterized in this way is <u>random</u>
- Example: Rolling of an ordinary six-sided dice.

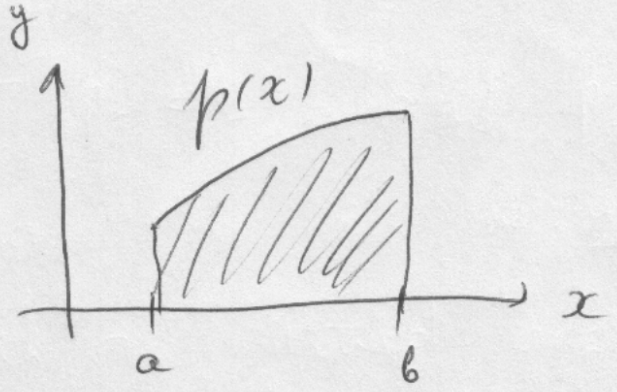
Measures of Data Distributions

- If the variable can take any real value, we have continuous random variable; and instead of p_i's, we have probability density function $p(x)$

$$p_i \Leftrightarrow p(x)\,dx$$

$$x = x_i \Leftrightarrow x \in [x, x+dx]$$





$$P(x \in [a,b]) = \int_a^b p(x)\,dx$$

- Measures of distributions:

$$\langle x^n \rangle = E(x^n) = \sum_{i=1}^{m} x_i^n p_i = \int_{-\infty}^{\infty} x^n p(x)\,dx$$

$$E(x) = \langle x \rangle = \bar{x} \leftarrow \text{Mean value}$$

Measures of Data Distributions

- How these moments are estimated using real (experimental) data?

$$E(x^n) = \langle x^n \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i^n$$

- Moments can be centered around some origin a:

$$E_a(x^n) = \frac{1}{N} \sum_{i=1}^{N} (x_i - a)^n$$

- Examples:

1) $a=0$, $n=1$ arithmetic mean value

2) $a=E_0(x)$, $n=2$, variance

3) $a=E_0(x)$, $n=3$, skew

4) $a=E_0(x)$, $n=4$, kurtosis (peakedness)

- other measures

1) Geometric mean

$$\sqrt[N]{x_1 \cdots x_N}$$

Measures of Data Distributions

2) Median : middle observation when data are arranged in order of size (or arithmetic mean of the middle two items if the number of data is even)

3) Mode : most frequent data

- Properties of moments

$$E_o(\alpha x + \beta y) = \alpha E_o(x) + \beta E_o(y)$$

$$x \geq 0 \Rightarrow E(x) \geq 0$$

- Variance

$$Var(x) = E_{\bar{x}}(x^2) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Measures possible deviations from the mean

- Why $E_{\bar{x}}(x)$ is not used? $E_{\bar{x}}(x) = 0$

Measures of Data Distributions

- Properties

$$\sigma_x^2 = Var(x) = E_{\bar{x}}(x^2) = E_0((x-\bar{x})^2) =$$

$$= E_0(x^2) - \left(E_0(x)\right)^2 = E_0(x^2) - \bar{x}^2 =$$

$$= \overline{x^2} - \bar{x}^2$$

- Standard deviation is $\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\overline{x^2} - \bar{x}^2}$

- Examples in Excel

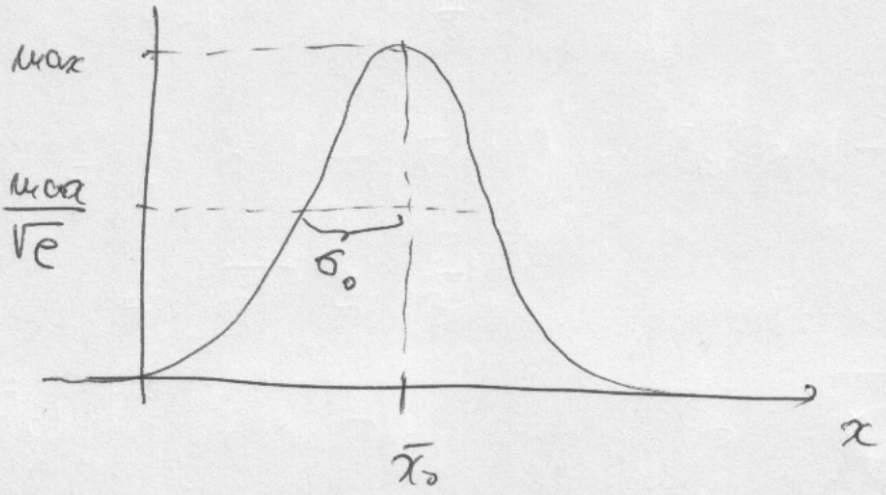
- Negative semi-variance and negative semi-deviation also used (just negative terms taken into account) — measure of financial risk

- We expect data to be distributed according to some distribution — often it is <u>normal</u> distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(x-\bar{x}_0)^2}$$

Measures of Data Distributions

- This distribution has $\sigma^2 = \sigma_0^2$, $\bar{x} = \bar{x}_0$



- If our empirical distribution is believed to be Gaussian, its parameters can be estimated as

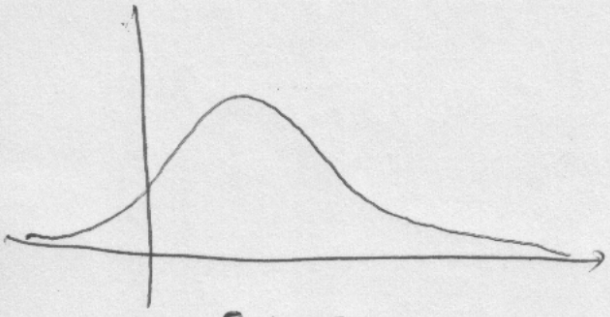
$$\bar{x}_0 = \bar{x}, \quad \sigma_0^2 = \sigma^2$$

- Skew and Kurtosis measure deviations from the normal distribution
- Skew measures asymmetry around the mean

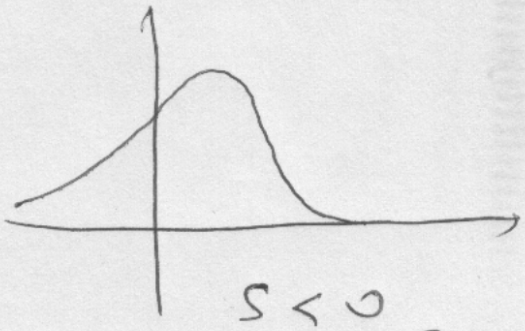
$$S_{kew} = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma_x} \right)^3$$

(7)

Measures of Data Distributing



$S > 0$



$S < 0$

$$S = \frac{1}{\sigma_x^3}\overline{(x-\bar{x})^3} \quad \text{for } N \text{ large}$$

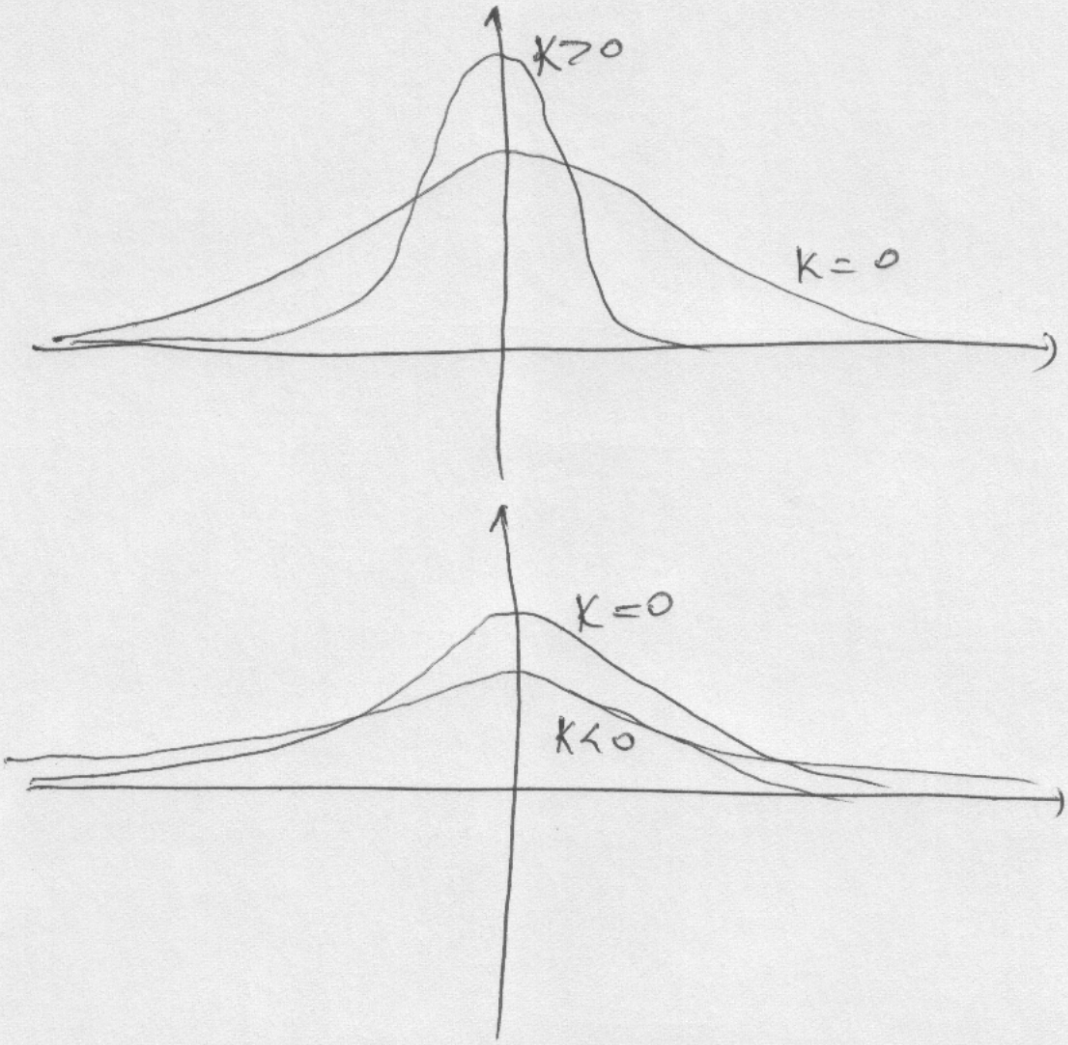
- Kurtosis characterizes the relative peakedness or flatness compared to the normal distribution

$$\text{Kurtosis} = \frac{N(N+1)}{(N-1)(N-2)(N-3)}\sum_{i=1}^{N} K_i - \frac{3(N-1)^2}{(N-2)(N-3)}$$

$$K_i = \frac{(x_i-\bar{x})^4}{\sigma_x^4}$$

For large N enough, $K = \frac{1}{\sigma_x^4}\overline{(x-\bar{x})^4} - 3$

Measures of Data Distributions



- When considering two or more random variables, their mutual dependance can be summarized by their covariance
- If we have two random variables, x and y, then

$$E(x) = \bar{x}, \quad E(y) = \bar{y},$$

$$\sigma_{x,y} = Cov(x,y) = E\left((x-\bar{x})(y-\bar{y})\right)$$

Measures of Data Distributions

- Covariance is

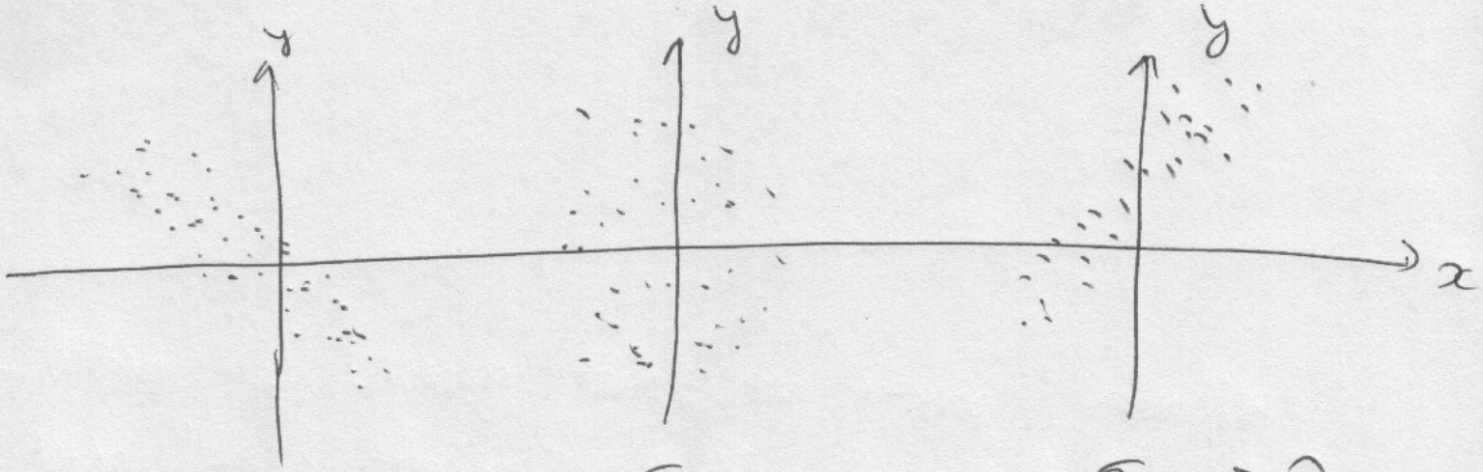
$$Cov(x,y) = E((x-\bar{x})(y-\bar{y})) = E(xy - x\bar{y} - \bar{x}y - \bar{x}\bar{y}) =$$

$$= E(xy) - \bar{x}\,\bar{y} = \overline{xy} - \bar{x}\,\bar{y}$$

- $Cov(x,x) = Var(x) = \sigma_x^2 = \sigma_{x,x}$

- Theorem states that $|\sigma_{x,y}| \leq \sigma_x \sigma_y$

- $\sigma_{x,y}$ desribes correlation of variables



$\sigma_{x,y} < 0$ $\quad \sigma_{x,y} = 0$ $\quad \sigma_{x,y} > 0$

$$Var(x+y) = Var(x) + 2\sigma_{x,y} + Var(y)$$

- If x and y are uncorrelated, $Var(x+y) = Var(x) + Var(y)$

(10)

Measures of Data Distributions

- Correlation matrix of several variables is made up out of covariances:

$$\begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix}$$

- It is symmetric, since $cov(x,y) = cov(y,x)$

- CFA 3 can be used at further reading

- HW1 Due next week!