Chapter 3, Part A
Descriptive Statistics: Numerical Measures(數值方法)

- Measures of Location
- Measures of Variability

Numerical Measures

- If the measures are computed for data from a sample, they are called sample statistics.
- If the measures are computed for data from a population, they are called population parameters.
- A sample statistic is referred to as the point estimator of the corresponding population parameter.

Measures of Location

- Mean
- Median
- Mode
- Weighted Mean
- Geometric Mean
- Percentiles
- Quartiles

Mean

- Perhaps the most important measure of location is the mean.
- The mean provides a measure of central location.
- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean μ.

Sample Mean $\bar{x}$

$$\bar{x} = \frac{\sum x_i}{n}$$

where: Σx_i = sum of the values of n observations
n = number of observations in the sample

Population Mean μ

$$\mu = \frac{\sum x_i}{N}$$

where: Σx_i = sum of the values of the N observations
N= number of observations in the population

Sample Mean $\bar{x}$

- Example: Apartment Rents

Seventy efficiency apartments were randomly sampled in a college town. The monthly rents for these apartments are listed below.

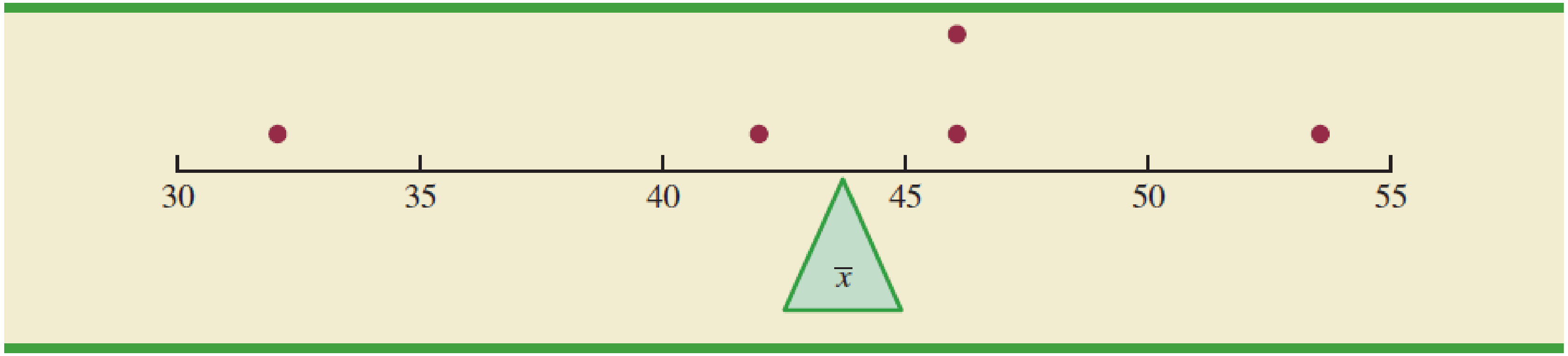
545	715	530	690	535	700	560	700	540	715
540	540	540	625	525	545	675	545	550	550
565	550	625	550	550	560	535	560	565	580
550	570	590	572	575	575	600	580	670	565
700	585	680	570	590	600	649	600	600	580
670	615	550	545	625	635	575	650	580	610
610	675	590	535	700	535	545	535	530	540

$$\bar{x} = \frac{\sum x_i}{n} = \frac{41{,}356}{70} = 590.80$$

79 87#75#79#65

$$\bar{x} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

FIGURE 3.1 THE MEAN AS THE CENTER OF BALANCE FOR THE DOT PLOT OF THE CLASSROOM SIZE DATA



79#447 75#79#65

$$\bar{x} = \frac{46 + \mathbf{114} + 42 + 46 + 32}{5} = \mathbf{56}$$

Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has **extreme values**, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median

- For an <u>odd number</u> of observations:

26	18	27	12	14	27	19	7 observations
12	14	18	19	26	27	27	in ascending order

The median is the middle value. Median = 19

Median

- For an <u>even number</u> of observations:

26	18	27	12	14	27	30	19	8 observations
12	14	18	19	26	27	27	30	in ascending order

The median is the average of the two middle values.

Median = (19 + 26)/2 = 22.5

Median

- Example: Apartment Rents

Averaging the 35th and 36th data values:

Median = (575 + 575)/2 = 575

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Note: Data is in ascending order.

Trimmed Mean

- Another measure, sometimes used when extreme values are present, is the trimmed mean.
- It is obtained by deleting a percentage of the smallest and largest values from a data set and then computing the mean of the remaining values.
- For example, the 5% trimmed mean is obtained by removing the smallest 5% and the largest 5% of the data values and then computing the mean of the remaining values.

Mode

- The <u>mode</u> of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are <u>bimodal</u>.
- If the data have more than two modes, the data are <u>multimodal</u>.

Mode

- Example: Apartment Rents

550 occurred most frequently (7 times)

Mode = 550

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Note: Data is in ascending order.

Weighted Mean

- In some instances the mean is computed by giving each observation a weight that reflects its relative importance.
- The choice of weights depends on the application.
- The weights might be the number of credit hours earned for each grade, as in GPA.
- In other weighted mean computations, quantities such as pounds, dollars, or volume are frequently used.

Weighted Mean

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where: x_i = value of observation i

w_i = weight for observation i

Numerator: sum of the weighted data values

Denominator: sum of the weights

If data is from a population, μ replaces $\bar{x}$.

Weighted Mean

- Example: Construction Wages

Ron Butler, a home builder, is looking over the expenses he incurred for a house he just built. For the purpose of pricing future projects, he would like to know the average wage ($/hour) he paid the workers he employed. Listed below are the categories of workers he employed, along with their respective wage and total hours worked.

Worker	Wage ($/hr)	Total Hours
Carpenter	21.60	520
Electrician	28.72	230
Laborer	11.80	410
Painter	19.75	270
Plumber	24.16	160

Weighted Mean

- Example: Construction Wages

Worker	x_i	w_i	$w_i x_i$
Carpenter	21.60	520	11232.0
Electrician	28.72	230	6605.6
Laborer	11.80	410	4838.0
Painter	19.75	270	5332.5
Plumber	24.16	160	3865.6
		1590	31873.7

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{31{,}873.7}{1{,}590} = 20.0464 = \$20.05$$

FYI, equally-weighted (simple) mean = $21.21

Geometric Mean

- The geometric mean is calculated by finding the *n*th root of the product of *n* values.
- It is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results).
- It should be applied anytime you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, . . .).
- Other common applications include: changes in populations of species, crop yields, pollution levels, and birth and death rates.

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\dots(x_n)}$$

$$= [(x_1)(x_2)\dots(x_n)]^{1/n}$$

Geometric Mean

- Example: Rate of Return

Period	Return (%)	(1 + Return%) Growth Factor
1	-6.0	0.940
2	-8.0	0.920
3	-4.0	0.960
4	2.0	1.020
5	5.4	1.054

$$\bar{x}_g = \sqrt[5]{(.94)(.92)(.96)(1.02)(1.054)}$$

$= [.89254]^{1/5} = .97752$

Average growth rate per period is (.97752 - 1) (100) = -2.248%

Exercise

Annual revenue for Corning Supplies grew by 5.5% in 2007; 1.1% in 2008; -3.5% in 2009; -1.1% in 2010; and 1.8% in 2011. What is the mean growth annual rate over this period?

TABLE 3.2 PERCENTAGE ANNUAL RETURNS AND GROWTH FACTORS FOR THE MUTUAL FUND DATA

Year	Return (%)	(1 + Return%) Growth Factor
1	−22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	−37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

$100 - .221($100) = $100(1 - .221) = $100(.779) = $77.90

$77.90 + .287(77.90) = 77.90(1 + .287) = $77.90(1.287) = $100.2573

$100(.779)(1.287) = $100.2573

$100[(.779)(1.287)···(1.021)] = $100(1.334493) = $133.4493

$$\bar{x}_g = \sqrt[10]{1.334493} = 1.029275$$

Average annual rate = (1.029275 – 1)100% = 2.9275%

Exercise

Suppose that at the beginning of 2004 you invested $10,000 in the Stivers mutual fund and $5,000 in the Trippi mutual fund. The value of each investment at the end of each subsequent year is provided in the table below. Which mutual fund performed better?

Year	Strivers	Trippi
2004	11,000	5,600
2005	12,000	6,300
2006	13,000	6,900
2007	14,000	7,600
2008	15,000	8,500
2009	16,000	9,200
2010	17,000	9,900
2011	18,000	10,600

Percentiles

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.
- The <u>*p*th percentile</u> of a data set is a value such that at least p percent of the items take on this value or less and at least (100 - p) percent of the items take on this value or more.
- Arrange the data in ascending order.
- Compute L_p, the location of the pth percentile.

$$L_p = (p/100)(n + 1)$$

80^{th} Percentile

- Example: Apartment Rents

$$L_p = (p/100)(n + 1) = (80/100)(70 + 1) = 56.8$$

(the 56^{th} value plus .8 times the difference between the 57^{th} and 56^{th} values)

80th Percentile = 635 + .8(649 – 635) = 646.2

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

80^{th} Percentile

- Example: Apartment Rents

"At least 80% of the items take on a value of 646.2 or less."

"At least 20% of the items take on a value of 646.2 or more."

56/70 = .8 or 80%

14/70 = .2 or 20%

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Quartiles

- Quartiles are specific percentiles.
- First Quartile (Q1) = 25th Percentile
- Second Quartile (Q2) = 50th Percentile = Median
- Third Quartile (Q3) = 75th Percentile

Third Quartile (75th Percentile)

- Example: Apartment Rents

$L_p = (p/100)(n + 1) = (75/100)(70 + 1) = 53.25$

(the 53rd value plus .25 times the difference between the 54th and 53rd values)

Third quartile = $625 + .25(625 - 625) = 625$

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

Measures of Variability

- Range
- Interquartile Range (四分位距)
- Variance +變異數,
- Standard Deviation
- Coefficient of Variation (變異係數)

Range

- The range of a data set is the difference between the largest and smallest data value.

Range = Largest value – Smallest value

- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Range

- Example: Apartment Rents

Range = largest value - smallest value

Range = 715 - 525 = 190

$\boxed{525}$	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	$\boxed{715}$

Interquartile Range (四分位距)

- The interquartile range of a data set is the difference between the third quartile and the first quartile (Q3－Q1).
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

Interquartile Range (IQR)

- Example: Apartment Rents

3rd Quartile ($Q3$) = 625

1st Quartile ($Q1$) = 545

IQR = $Q3$ - $Q1$ = 625 - 545 = 80

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Variance（變異數）

- The variance is a measure of variability that utilizes all the data.
- It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population).
- The variance is useful in comparing the variability of two or more variables.
- The variance is the average of the squared deviations between each data value and the mean.
- The variance is computed as follows:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \qquad \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

for a sample   for a population

Degrees of Freedom

- The concept of degrees of freedom is central to the principle of estimating statistics of populations from samples of them. "Degrees of freedom" is commonly abbreviated to df.
- The number of degrees of freedom generally refers to the number of independent observations in a sample minus the number of population parameters that must be estimated from sample data.
- The degrees of freedom of an estimate is the number of independent pieces of information on which the estimate is based.

TABLE 3.3 COMPUTATION OF DEVIATIONS AND SQUARED DEVIATIONS ABOUT THE MEAN FOR THE CLASS SIZE DATA

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean ($(x_i - \bar{x})^2$)
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		$\boxed{0}$	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$ +vwxghqw,5

TABLE 3.4 COMPUTATION OF THE SAMPLE VARIANCE FOR THE STARTING SALARY DATA

Monthly Salary (x_i)	Sample Mean ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
3850	3940	−90	8,100
3950	3940	10	100
4050	3940	110	12,100
3880	3940	−60	3,600
3755	3940	−185	34,225
3710	3940	−230	52,900
3890	3940	−50	2,500
4130	3940	190	36,100
3940	3940	0	0
4325	3940	385	148,225
3920	3940	−20	400
3880	3940	−60	3,600
		0	301,850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Using equation (3.7),

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301,850}{11} = 27,440.91$$

Standard Deviation (標準差)

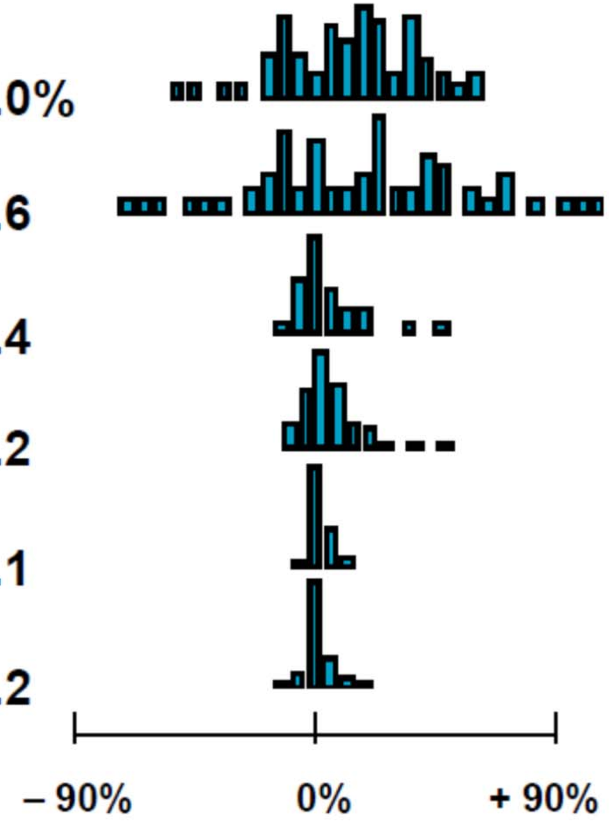
- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation is computed as follows:

$$s = \sqrt{s^2} \qquad\qquad \sigma = \sqrt{\sigma^2}$$

for a sample　　　　for a population

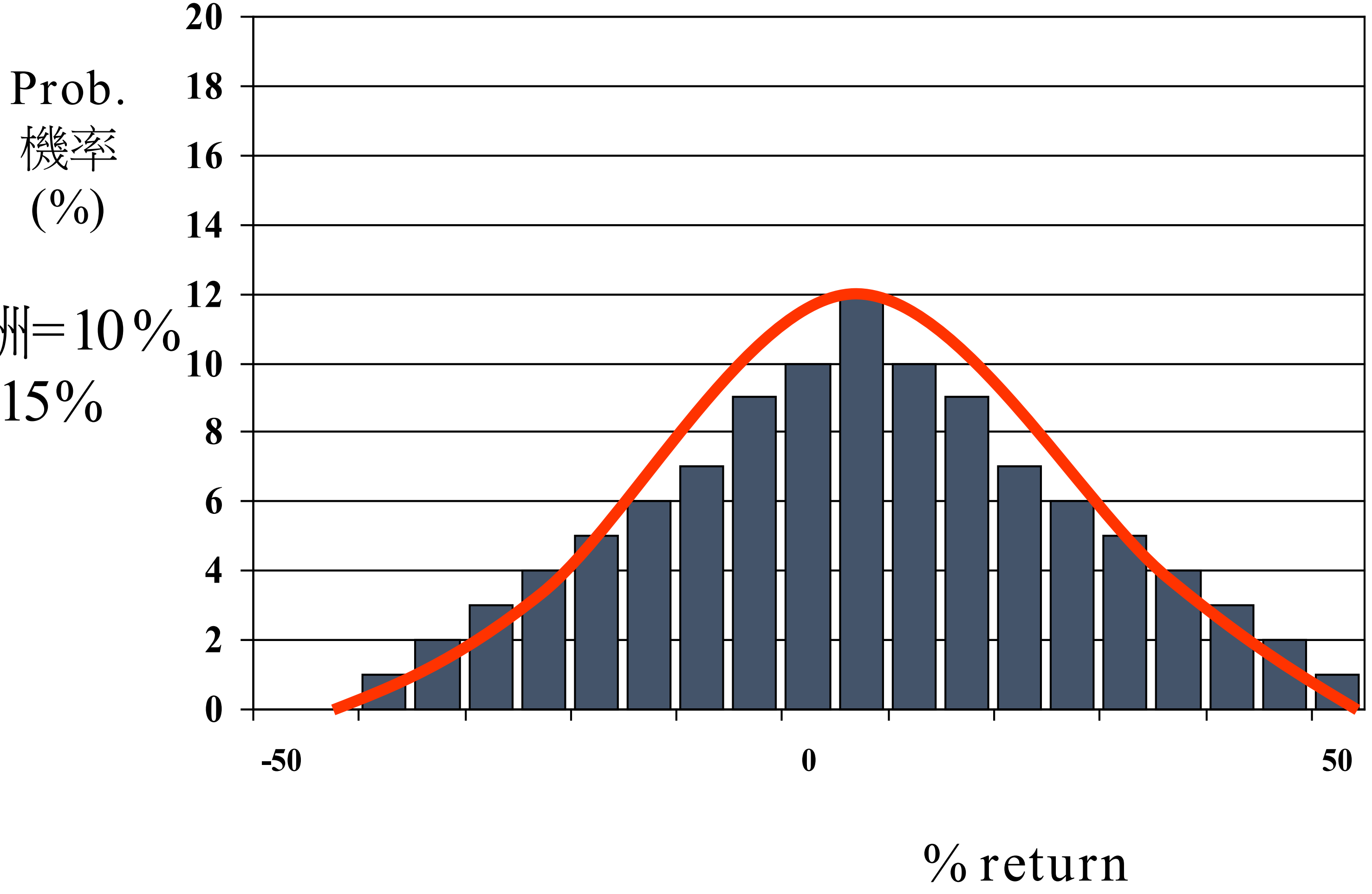
歷史報酬: 1926-2007

類別	平均年報酬率	標準差	分布
Large Company Stocks	12.3%	20.0%	
Small Company Stocks	17.1	32.6	
Long-Term Corporate Bonds	6.2	8.4	
Long-Term Government Bonds	5.8	9.2	
U.S. Treasury Bills	3.8	3.1	
Inflation	3.1	4.2	



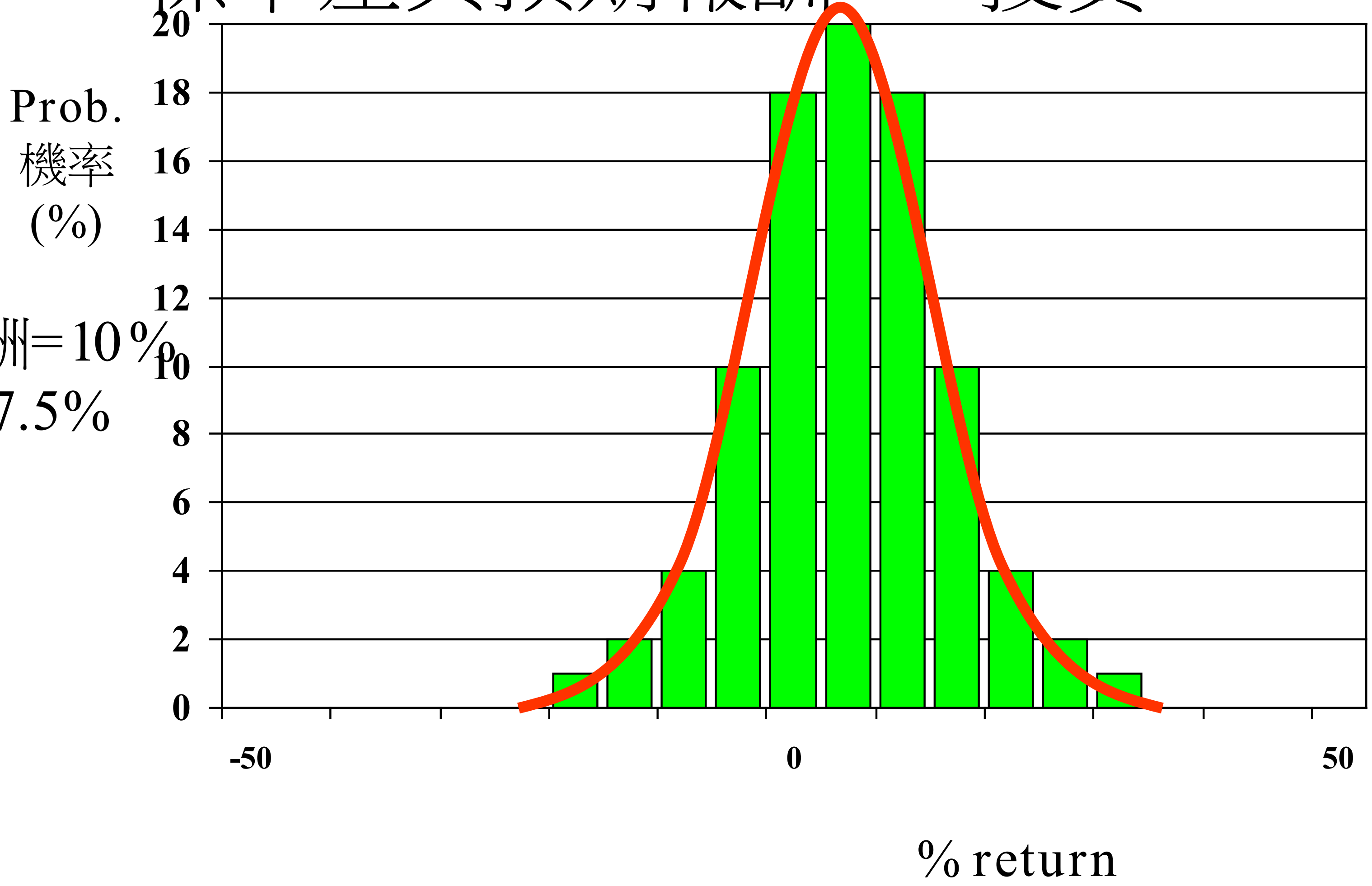
Standard deviation and expected returns: Investment A
標準差與預期報酬：A投資

Avg. ret. 平均報酬=10%
STD 標準差 =15%

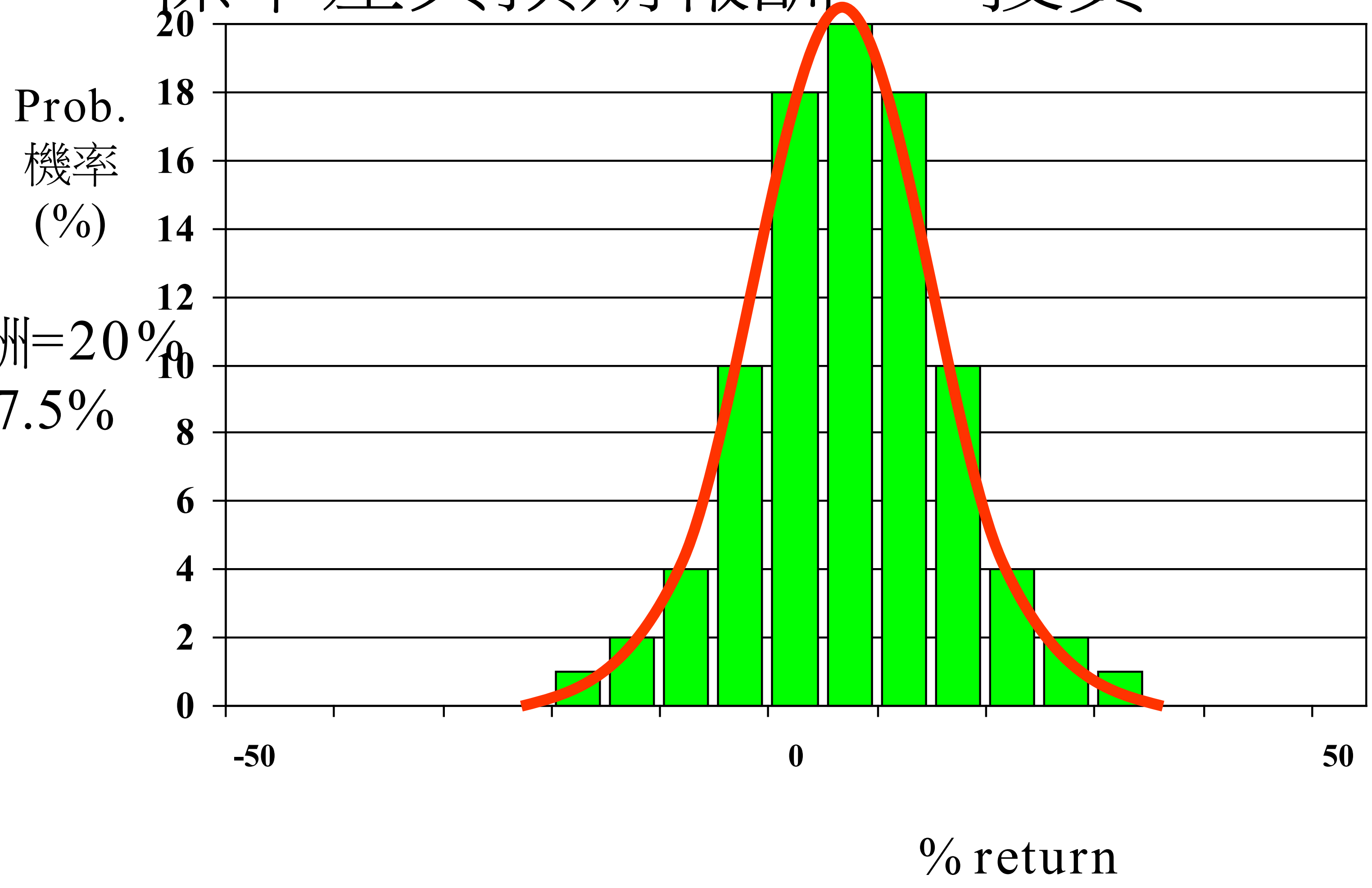


Standard deviation and expected returns: Investment B
標準差與預期報酬：B投資

Avg. ret. 平均報酬=10%
STD 標準差 =7.5%



Standard deviation and expected returns: Investment C
標準差與預期報酬：C投資



Avg. ret. 平均報酬=20%
STD 標準差 =7.5%

Coefficient of Variation (變異係數)

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.

- The coefficient of variation is computed as follows:

$$\left[\frac{s}{\bar{x}} \times 100\right]\% \qquad \left[\frac{\sigma}{\mu} \times 100\right]\%$$

for a sample	for a population

Sample Variance, Standard Deviation, And Coefficient of Variation

- Example: Apartment Rents
  - Variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = 2{,}996.16$$

  - Standard Deviation

$$s = \sqrt{s^2} = \sqrt{2{,}996.16} = 54.74$$

  - Coefficient of Variation

$$\left[\frac{s}{\bar{x}} \times 100\right]\% = \left[\frac{54.74}{590.80} \text{ x } 100\right]\% = 9.27\%$$

Mean absolute error (平均絕對誤差，MAE)

$$MAE = \frac{\sum|x_i - \overline{x}|}{n}$$

Class size data: $MAE = \frac{28}{5} = 5.6$