



COLUMBIA  
BUSINESS  
SCHOOL

## Descriptive Statistics

Fall 2001  
B6014: Managerial Statistics

Professor Paul Glasserman  
403 Uris Hall

### Histograms

1. A *histogram* is a graphical display of data showing the frequency of occurrence of particular values or ranges of values. In a histogram, the horizontal axis is divided into bins, representing possible data values or ranges. The vertical axis represents the number (or proportion) of observations falling in each bin. A bar is drawn in each bin to indicate the number (or proportion) of observations corresponding to that bin. You have probably seen histograms used, e.g., to illustrate the distribution of scores on an exam.
2. All histograms are bar graphs, but not all bar graphs are histograms. For example, we might display average starting salaries by functional area in a bar graph, but such a figure would not be a histogram. Why not? Because the Y-axis values do not represent relative frequencies or proportions, and the X-axis values do not represent value ranges (in particular, the order of the bins is irrelevant).

### Measures of Central Tendency

1. Let  $X_1, \dots, X_n$  be data points, such as the result of  $n$  measurements or observations. What one number best characterizes or summarizes these values? The answer depends on the context. Some familiar summary statistics are these:
  - The *mean* is given by the arithmetic average  $\bar{X} = (X_1 + \dots + X_n)/n$ . (Notation: We will often write

$$\sum_{i=1}^n X_i \quad \text{for} \quad X_1 + \dots + X_n.$$

The symbol  $\sum_{i=1}^n X_i$  is read “the sum from  $i$  equals 1 up to  $n$  of  $X_i$ .”)

- The *median* is larger than one half of the observations and smaller than the other half. If the number of data points is odd, then the median is the middle value; if the number of data points is even, then the median is the average of the two middle values.
  - The *mode* is the value that appears with the greatest frequency. There may be multiple modes.
2. The following statements give some information about the “typical” American based on the 2000 census. These statements use the word “average” in a vague way: in everyday usage, “average” doesn’t always signify “mean.” For each statement, think about which summary statistic is probably behind the information reported.
- The average American is a white woman, 35.4 years old.
  - The average household size for owner-occupied units is 2.71 people.
  - The average American travels 24.3 minutes to work and works in the service sector.
  - The median household income (including benefits) is \$41,343; the mean is \$55,253.
  - The median income for males is \$30,132; for females, \$18,996.
  - The average consumer credit card debt is \$5,451.

Careful consideration of these examples suggests that the mode is useful in summarizing *categorical* (i.e., non-numeric) data, to which the other measures are inapplicable. The sense in which the typical American is a woman is that there are more women than men.

3. When are the mean and median equal? When is the mean larger? When is the median larger? The mean and median are equal whenever the distribution of observations (as depicted by a histogram) is symmetric. The mean is larger (smaller) if the distribution is skew right (left). See Figure 1.
4. Suppose for simplicity that the data points  $X_1, \dots, X_n$  can only take the values  $1, 2, 3, \dots$ . For  $i = 1, 2, 3, \dots$ , let  $p(i)$  be the *proportion* of data points equal to  $i$ . The mean, defined above, is also given by

$$\bar{X} = p(1) \cdot 1 + p(2) \cdot 2 + p(3) \cdot 3 + \dots \quad (1)$$

In this expression, each possible value is weighted by its frequency of occurrence. This formula is particularly convenient when the data appears in a histogram: it weights each bin-value by the (relative) height of the bar in that bin.

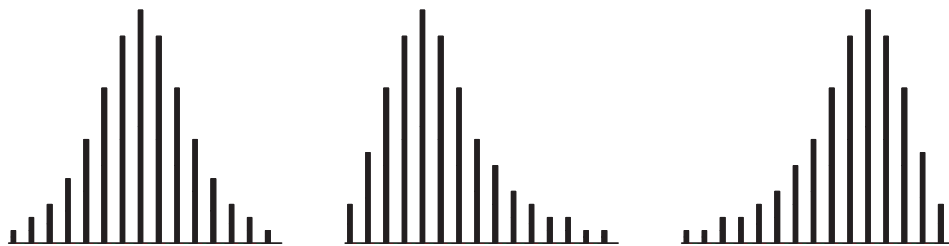


Figure 1: Examples of distributions that are symmetric, skew right, and skew left.

5. Another way to look at the mean is to think of giving equal weight  $1/n$  to each observation  $X_i$ . More generally, if  $w_1, \dots, w_n$  are positive numbers summing to 1, the **weighted average** determined by these numbers is

$$w_1X_1 + \dots + w_nX_n = \sum_{i=1}^n w_iX_i.$$

6. **Stock indices** are an important class of summary statistics. The simplest way to define an index is to pick  $n$  stocks with prices  $X_1, \dots, X_n$  and let the index be the mean price. (In contrast to an ordinary average, an index is usually defined relative to a base value.)

What happens to this index if one or more of the  $n$  stocks splits (e.g., if its number of shares is doubled and the price is cut in half)? What does this say about the index?

The **Dow Jones Industrial Average** compensates for splits by adjusting the denominator: there are 30 stocks in this index, but the sum of their prices is not divided by 30; it is divided by a “divisor” that changes over time. Each time a stock splits, the denominator is lowered so that the index is unaffected by the split. For example, suppose you had an index of stocks with prices \$5, \$10, and \$15 for an average of \$10. If the \$15 stock splits 3-for-1, each new share should be worth \$5 so if we didn’t adjust the divisor the average would drop to  $(5 + 10 + 5)/3 = 6.67$ . Instead, to compensate for the split we change the divisor to 2 so that the index stays at  $(5 + 10 + 5)/2 = 10$ . In August 2001, the average price of the Dow 30 was about \$50. The divisor was at 0.14452124. If we divide the sum of the prices of the Dow 30 by the divisor rather than by 30 we get the index level (around 10,400, rather than the average price of about 50). For further explanation of these calculations, visit [www.dowjones.com](http://www.dowjones.com).

7. Another way to get around the problem of splits in a stock index is to weight each stock by the number of shares: if there are  $w_i$  shares of stock  $i$ , then the index is

$$\frac{1}{c} \sum_{i=1}^n w_iX_i,$$

where  $c$  is a fixed proportionality constant. If a stock in the index splits, the number of shares doubles but the price is halved so the product  $w_i X_i$  is unchanged. The index is thus unaffected by splits. The **S&P 500 Index** is of this type. There, the constant  $c$  is ten times the average value of the numerator in a base period, 1941-1943. A criticism of this type of capitalization-weighted index is that it is dominated by a small number of large companies.

8. Suppose you earn interest at the monthly rates  $r_1, r_2, \dots, r_{12}$  from January to December. Your total compounded interest is the same as it would have been at a fixed monthly rate  $r$ , for some  $r$ . This  $r$  is, in some sense, the “average” of the twelve monthly rates. How do you find  $r$ ? Answer:

$$r = [(1 + r_1)(1 + r_2) \cdots (1 + r_{12})]^{1/12} - 1.$$

What would the equivalent fixed rate be without compounding? Answer:

$$r = (r_1 + r_2 + \cdots + r_{12})/12.$$

9. The *geometric mean* of positive numbers  $X_1, \dots, X_n$  is given by

$$(X_1 X_2 \cdots X_n)^{1/n},$$

the  $n$ -th root of the product of the  $n$  values. Thus, the geometric mean gives the equivalent fixed monthly rate from twelve different monthly rates when interest is compounded. This is useful, but we won't be using the geometric mean in this course.

## Measures of Variability

1. Once a reasonable measure of central tendency has been identified, the next step is to summarize the extent of deviation from this central tendency or to measure the degree of variability in the data.
2. The crudest measure of spread in a dataset is its **range**: the difference between the largest and smallest value. But the minimum and maximum values of a dataset are very sensitive to outliers, so the range is similarly sensitive.
3. The **interquartile range** is the length of the interval containing the middle 50% of the data; i.e., the spread from the 25<sup>th</sup> percentile<sup>1</sup> to the 75<sup>th</sup> percentile. The interquartile range supplements the median. It is less sensitive to outliers than the range.

---

<sup>1</sup>Percentiles are easiest to define by example: 25% of the observations fall below the 25<sup>th</sup> percentile, 75% below the 75<sup>th</sup> percentile, 95% below the 95<sup>th</sup> percentile, etc.

4. The **mean absolute deviation**, or MAD, is given by

$$\frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|. \quad (2)$$

In words, the MAD is the average distance of the data points from the mean. The mean absolute deviation is minimized when the mean is replaced by the median; that is, if we could replace  $\bar{X}$  in (2) with another value, the value that would minimize the expression is the median. In this sense, the median is the point that best balances absolute errors.

5. The most important measure of spread in a dataset is the **standard deviation**, denoted by  $\sigma$ . The standard deviation is calculated as the square root of the **variance**  $\sigma^2$ , which is given by

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (3)$$

Thus, the standard deviation is

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (4)$$

In practice, it is usually easier to use the following alternative formula:

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

and then take the square root to get  $\sigma$ .

6. The  $\sigma$  and  $\sigma^2$  defined above are the so-called *population* standard deviation and variance. They assume that the entire population of data consists of our  $n$  data points. If our  $n$  points are merely a sample of a larger population, then we would divide by  $n - 1$  rather than  $n$  to get the *sample* standard deviation and variance. We don't need to worry about this distinction until later.
7. The mean  $\bar{X}$  minimizes the sum of squared errors in (3); i.e., replacing  $\bar{X}$  in that expression with any value can only increase the result. In this sense, the mean balances *squared* deviations.
8. One of the main themes of this course is that the standard deviation is the fundamental unit of distance in a statistical setting. Here are some initial indications of what this means:
- (a) Is an observation unusually high or low? Look at how many standard deviations away from the mean it is.

- (b) Is a fluctuation in a stock price unusually large? Look at how many standard deviations the price moved.
- (c) Do Finance concentrators have significantly higher average starting salaries than Accounting concentrators? Compare the difference in means with the standard deviations.
9. Just by knowing the mean and standard deviation of a dataset, it is possible to say quite a bit about the possible spread of the values, using **Chebyshev's rule**: The fraction of data points within  $m$  standard deviations of the mean, for any  $m > 1$ , is at least  $1 - (1/m^2)$ . Thus, at least
- 55.6% lie within 1.5 standard deviations of the mean;
  - 75% lie within 2;
  - 88.9% lie within 3.

This is true for *all* data sets.

10. **EXAMPLE.** You were sick on the day of the midterm. Before making up the exam, you learn that the average score was 73 and the standard deviation was 7. Suppose the top 25% of the class earns an H. According to Chebyshev, what score will guarantee you an H?

From Chebyshev's rule, we know that at least 75% of scores are within 2 standard deviations of the mean, so at most 25% of scores can be more than 2 standard deviations above the mean. So, a score 2 standard deviations above the mean will guarantee an H. This score is  $73 + 2 \times 7 = 87$ .

What if only the top 20% get an H? How many standard deviations above the mean do you need to score now? Let  $m$  be the required number of standard deviations above the mean. Chebyshev's rule tells us to choose  $m$  so that  $1/m^2 = 0.20$ ; i.e.,  $m = 2.236$ . The score that guarantees an H is now  $73 + (2.236) \times 7 \approx 88.7$ .

11. Since Chebyshev's inequality holds for all datasets, it is somewhat pessimistic; typically, values will be more clustered around the mean. A frequently used rule of thumb for large populations is that approximately 68% of the observations lie within one standard deviation of the mean, 95% lie within two standard deviations. This rule of thumb is based on the **normal distribution** which we will study later in the course. In practice, this rule is used much more often than Chebyshev.

## Measures of Shape

1. Throughout much of this course we will be assuming that we are working with normally distributed data—i.e., data whose histogram follows a bell-shaped curve. It is worth considering, however, ways in which data can deviate from this assumption.

2. Perhaps the most dramatic way is that data can be *multimodal*, meaning that its histogram can have many peaks. In particular, a distribution is *bimodal* if it has two peaks. Often, multiple peaks reflect distinct underlying populations. For example, the distribution of heights of adults is probably bimodal, with one peak for the average height of women and one for the average height of men. As in this example, it is often possible to separate a multimodal distribution into separate unimodal populations.
3. *Skew* is another attribute that can distinguish data from a bell-shaped curved. The Excel function SKEW applied to values  $X_1, \dots, X_n$  produces

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^3.$$

Values much larger than the mean will tend to make this expression positive whereas values much smaller than the mean will tend to make it negative. Thus, positive values of skew indicate a distribution that is skew right and negative values of skew indicate a distribution that is skew left. (Positive skew, according to this definition, is not exactly the same as the condition that the mean be greater than the median, but the two notions of skew will usually agree if the asymmetry in the distribution is very pronounced. The same holds for negative skew.)

4. A symmetric, unimodal distribution may differ from a normal distribution by being excessively or insufficiently peaked. Peakedness is usually measured by *kurtosis*. A distribution with high kurtosis is very peaked.
5. If we compare two symmetric distributions with the same mean and the same standard deviation, the one with higher kurtosis will have more observations near the mean. It will also have more observations in the extreme tails—it has to for the standard deviations to work out to be the same. This situation is illustrated in Figure 2. The histogram has higher kurtosis than the bell-shaped curve; it has more observations near the mean and in the extremes of the tails.
6. High kurtosis is characteristic of the distribution of returns of financial data, particularly over short time periods—e.g., daily returns. Most days produce small changes but occasionally a price will make a sudden sharp move. The combination results in high kurtosis.
7. There is a minor ambiguity concerning the numerical scale on which kurtosis is measured. On the usual scale, normally distributed data has a kurtosis of three. The Excel function KURT measure *excess* kurtosis, meaning kurtosis beyond that of normal data. So, KURT would return a value of zero (rather than three) on normal data. This distinction does not really present a problem because we are interested in kurtosis only as a comparative measure. Just remember, higher kurtosis means more peaked and more observations in the extreme tails.

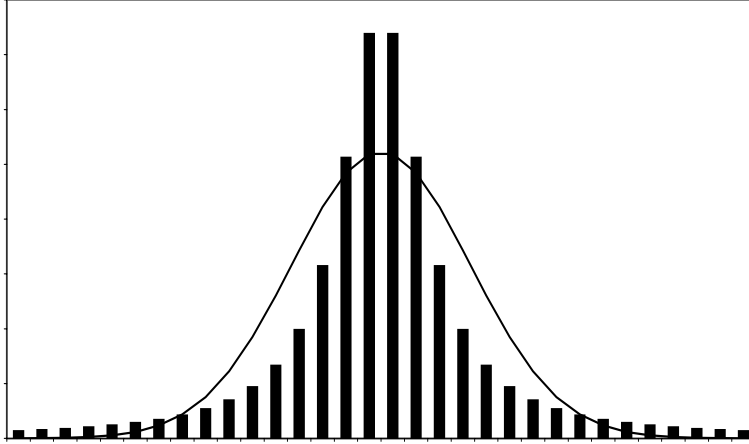


Figure 2: The distribution represented by the bars has greater kurtosis than the bell-shaped curve.

### Measures of Association

1. Thus far, we have discussed statistics describing only a single set of observations. We now turn to measures of relationships between two sets of observations. In general, these types of measures try to quantify the information about one variable contained in another, and to answer the following types of questions:
  - What does knowing a person’s height tell us about his or her likely weight?
  - What does knowing last week’s return on IBM stock tell us about next week’s likely return?
  - What does knowing the result of test marketing in Cleveland tell us about likely sales in Sacramento?
2. Suppose we have a set of pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$  of data points (e.g.,  $X_i$  is the  $i$ th person’s height and  $Y_i$  is that person’s weight). The most important measure of the relationship between the  $X$  values and the  $Y$  values is their *correlation*. To define correlation, we first need to introduce the **covariance**

$$Cov[X, Y] = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

The covariance looks a bit like the variance; in fact,  $Cov[X, X]$  is just the variance of  $X$ .

3. When is  $Cov[X, Y]$  large? It’s large when the individual terms  $(X_i - \bar{X})(Y_i - \bar{Y})$  tend to be large. When are these terms large? They are large if  $Y$  tends to be above its mean when  $X$  is above its mean. More precisely, they are large if the



values of  $i$  for which  $X_i > \bar{X}$  tend to coincide with the values of  $i$  for which  $Y_i > \bar{Y}$ . Thus, when you see a large, positive covariance think, “Above-average values of  $Y$  tend to occur together with above-average values of  $X$ .”

4. By essentially the same argument, a large *negative* covariance has the opposite interpretation: “Above-average values of  $Y$  tend to occur together with *below*-average values of  $X$ .” If the covariance is not very large in either the positive or negative direction, then the relation between  $X$  and  $Y$  is not very strong.
5. The inevitable next question is, “How large is large, how strong is strong?” A shortcoming of covariance as a measure of association is that its magnitude depends on the units in which the data is measured and so has no intrinsic meaning. For example, the covariance between height and weight, with height measured in inches is 12 times as large as the covariance with height measured in feet. So, it is impossible to interpret the magnitude of a covariance without knowing the context in which it is measured.
6. The **correlation coefficient** addresses this shortcoming by measuring the relationship between two variables on a fixed scale. By definition, we have

$$\text{Corr}[X, Y] = \rho_{XY} = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y},$$

where  $\sigma_X$  and  $\sigma_Y$  are the standard deviations of  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$ , respectively. We usually use  $\rho$ , the Greek letter rho, to denote a correlation coefficient.

7. Correlation is always between  $-1$  and  $1$ . A correlation of  $1$  reflects the strongest possible *positive* dependence between two variables, and a correlation of  $-1$  reflects the strongest possible *negative* dependence between variables. A correlation near  $0$  reflects very weak dependence (if any) between the two variables.
8. In the preceding paragraph, we described  $\rho$  as a measure of dependence. More precisely, correlation is a measure of *linear* dependence: the tendency of the  $(X_i, Y_i)$  pairs to fall on a straight line. In fact,  $\rho_{XY} = 1$  if and only if there are constants  $a, b$ , with  $b > 0$ , for which

$$Y_i = a + bX_i, \quad i = 1, \dots, n.$$

In words,  $\rho_{XY} = 1$  corresponds to a situation in which the  $(X_i, Y_i)$  pairs lie exactly on a straight line with positive slope. Similarly,  $\rho_{XY} = -1$  corresponds to a situation in which they lie on a straight line with *negative* slope  $b < 0$ . A value of  $\rho$  close to zero indicates that the  $(X_i, Y_i)$  pairs do not tend to fall on a line, and look more like a cloud.

9. If the absolute value  $|\rho_{XY}|$  is large (meaning close to  $1$ ), then information about  $X$  is useful in predicting the outcome of  $Y$ . If  $|\rho_{XY}|$  is close to zero, then information about  $X$  is not very useful in predicting  $Y$ .

10. WARNING: The magnitude of  $\rho$  reflects the *strength* of a linear relationship, not the *steepness* of the relationship (as measured by the slope  $b$ ). The points  $(X_i, Y_i)$  could be tightly clustered around a shallow line (large  $|\rho|$ , small  $|b|$ ) or loosely grouped around a steep line (small  $|\rho|$ , large  $|b|$ ).
11. Let's consider an example. Below are the first 11 rows of a table of 228 assessed property values in three Long Island towns. Each row corresponds to a single house. The numbers under VALUE are the property values. Under LOC, a 1 indicates East Meadow, a 4 indicates Levittown, and a 3 indicates Islip. Lot size is measured in thousands of square feet. The remaining variables should be self-explanatory.

ROW	VALUE	LOC	LOTSZ	BDRM	BATH	ROOMS	AGE	GARG
1	190.00	3	6.90	4	2.0	8	38	1
2	215.00	1	6.00	2	2.0	7	30	1
3	160.00	3	6.00	3	2.0	6	35	0
4	195.00	1	6.00	5	2.0	8	35	1
5	163.00	3	7.00	3	1.0	6	39	1
6	159.90	3	6.00	4	1.0	7	38	1
7	160.00	1	6.00	2	1.0	7	35	1
8	195.00	3	6.00	3	2.0	7	38	1
9	165.00	3	9.00	4	1.0	6	32	1
10	180.00	3	11.20	4	1.0	9	32	1
11	181.00	3	6.00	5	2.0	10	35	0
. . .								

To what extent does each of the property attributes *explain* or help predict property values?

12. One way to explore relationships among these variables is to look at the matrix of correlation coefficients:

	VALUE	LOTSZ	BDRM	BATH	ROOMS	AGE	GARG
VALUE	1						
LOTSZ	-0.059	1					
BDRM	0.209	0.150	1				
BATH	0.540	0.109	0.275	1			
ROOMS	0.481	0.171	0.525	0.466	1		
AGE	-0.252	-0.185	-0.142	-0.161	-0.152	1	
GARG	0.233	0.127	0.048	0.260	0.236	-0.052	1

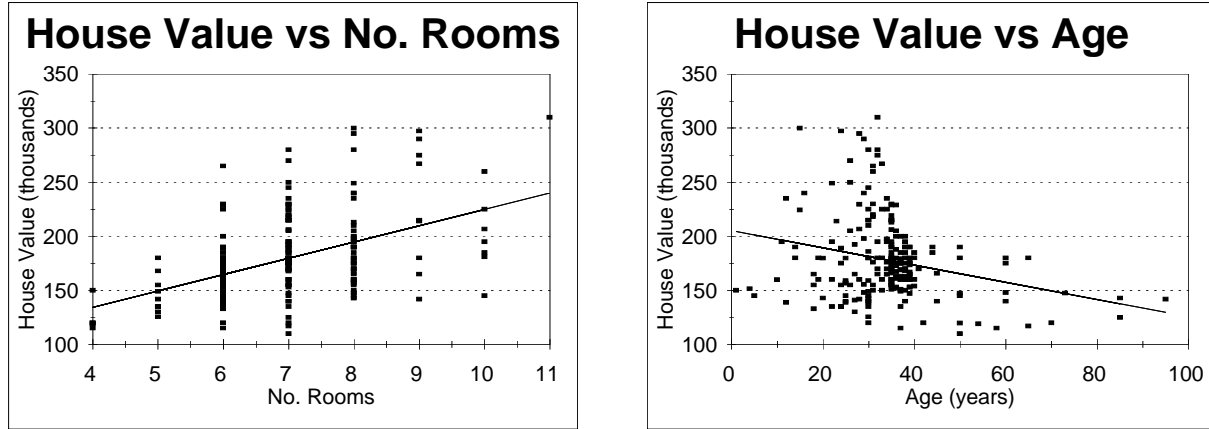


Figure 3: Scatter plots of house value against number of rooms and age

Each variable is perfectly correlated with itself, so we have 1's on the diagonal. The upper half of the table is omitted because the  $Corr(X, Y) = Corr(Y, X)$ , so the upper half is redundant.

13. Not surprisingly, the number of rooms is positively correlated with the value of the house, and the relationship is fairly strong. The age of a house is negatively correlated with its value, but that relationship is not as strong.
14. These relationships are further examined in Figure 1, where we have scatter plots relating house values to rooms and to age. (Each dot in each scatter plot corresponds to a pair of attributes for a single house.) The scatter plots are consistent with the positive and negative relations noted above. They are also consistent with the strengths noted above: there appears to be somewhat more of a linear relationship in the first scatter plot than the second.
15. The line drawn through each of the scatter plots is, in a precise sense, the best linear fit to the data. This is called a **regression line**. (We will discuss the precise sense in which this is the best fit later in the course.) The regression line helps summarize the relationship between pairs of variables. Its effectiveness in doing so depends on how strongly the variables are correlated.
16. As an example, suppose we ask, “What is the average value of houses with 5 rooms?” If we find 5 on the X-axis of the first graph and go up to the regression line, we find the value 150,000.
17. More importantly, suppose we want the average value of houses 47 years old. It turns out that there is not a single 47-year old house among the 228 in our table. Nevertheless, from the regression line we can predict that the average value of such houses (drawn from the same population as those in our table) should be about 168,000.

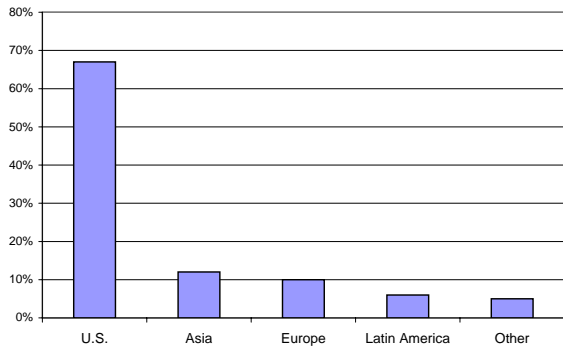
18. Later in the course, we will see how to do this kind of analysis more precisely. We will also see how to incorporate multiple variables simultaneously, instead of just two at a time. This will allow us, for example, to separate the value of additional rooms from the value of additional lot size.
19. We briefly mention how the regression line is determined (though we don't really need to know this now). The slope, typically denoted by the Greek letter  $\beta$ , is given by

$$\beta = \frac{Cov[X, Y]}{\sigma_X^2}.$$

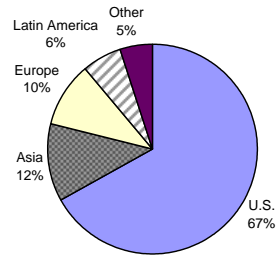
The intercept is then determined by the requirement that the line pass through the point  $(\bar{X}, \bar{Y})$ .

20. Final remark: You may have heard of a stock's *beta*. This is a special case of the  $\beta$  above, in which  $Y$  denotes returns on the stock and  $X$  denotes returns on the entire market. A stock's beta is a measure of how sensitive its price is to movements in the market as a whole.

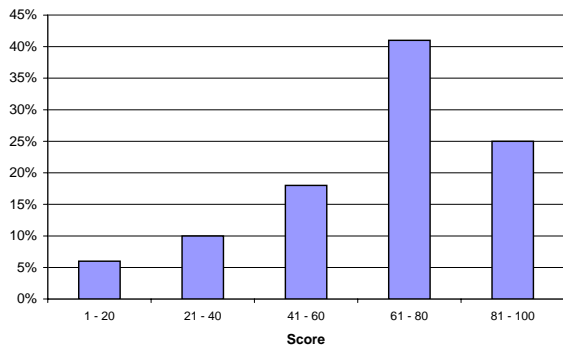
1998 MBA Class: Region of Origin



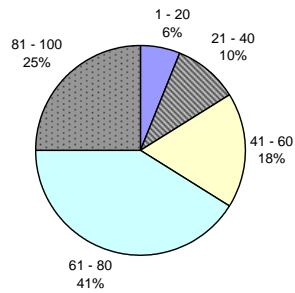
1998 MBA Class: Region of Origin



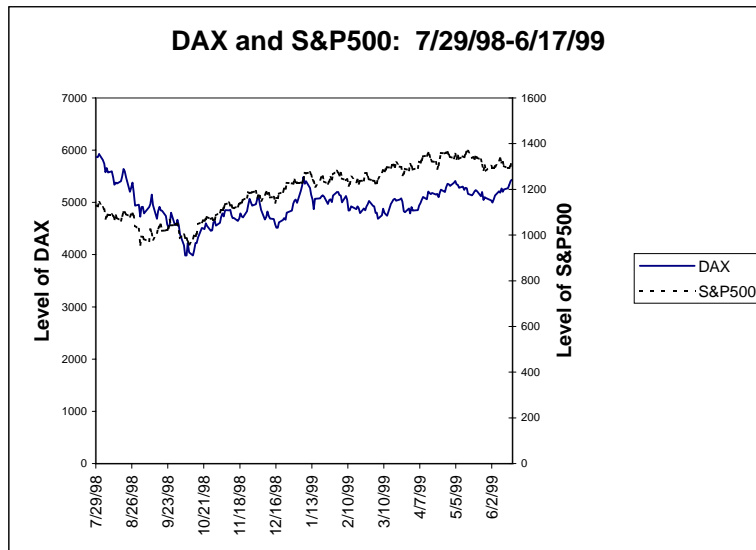
Exam Scores (124 Students)

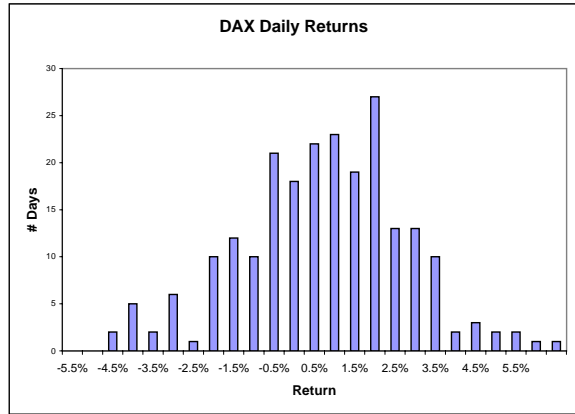
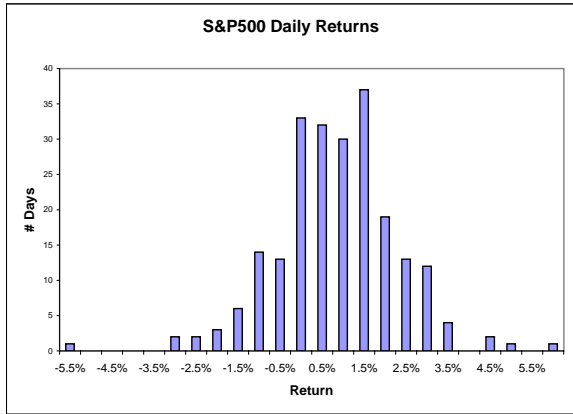


Exam Scores (124 Students)

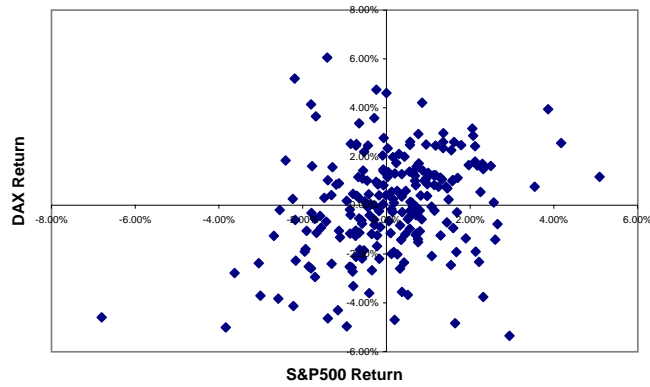


DAX and S&P500: 7/29/98-6/17/99





**DAX vs S&P500 Daily Returns  
(7/29/98-6/16/99)**



**NASDAQ vs S&P500  
Jan-Dec 98**

