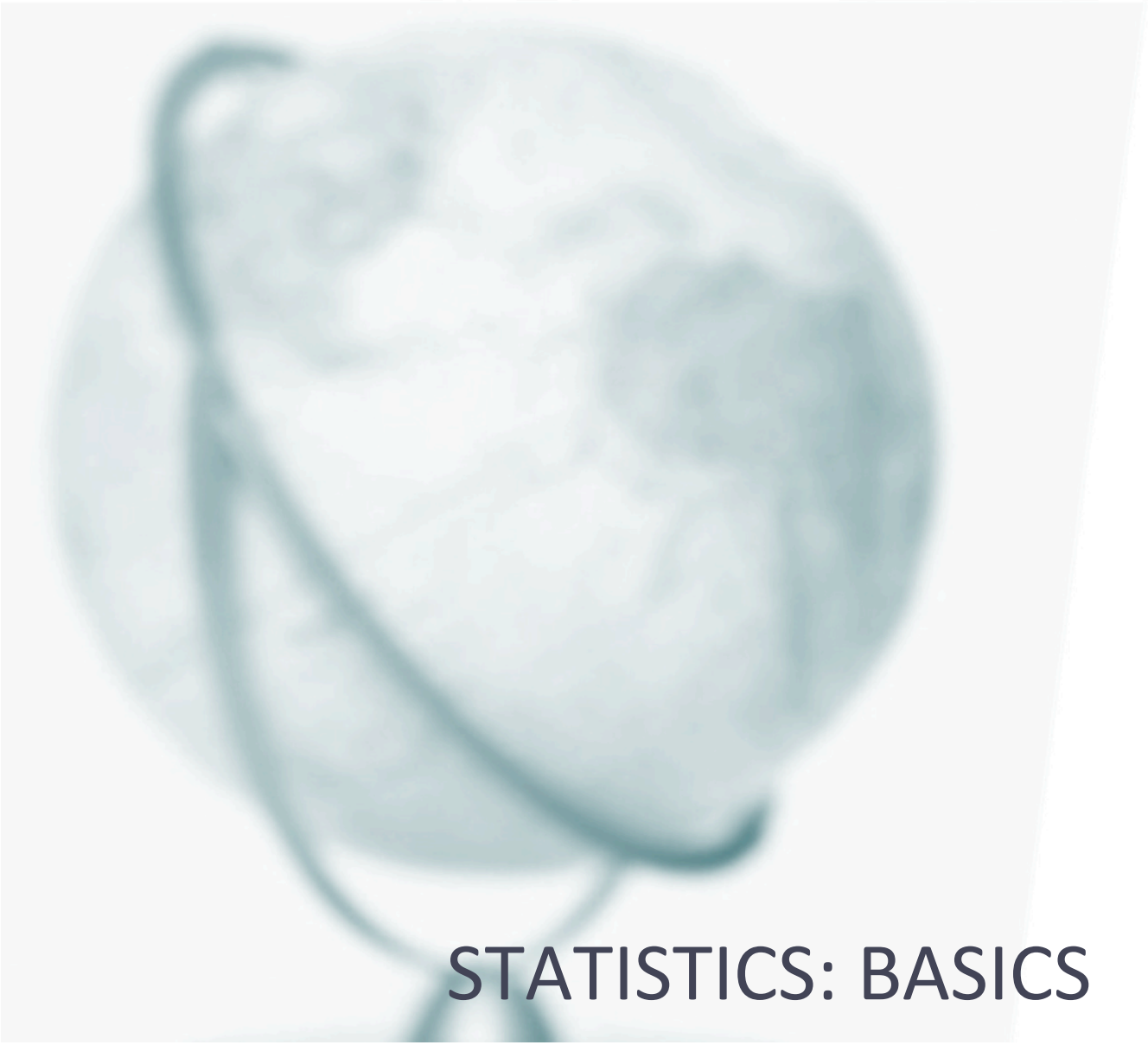


STATISTICS: BASICS

Aswath Damodaran

The role of statistics

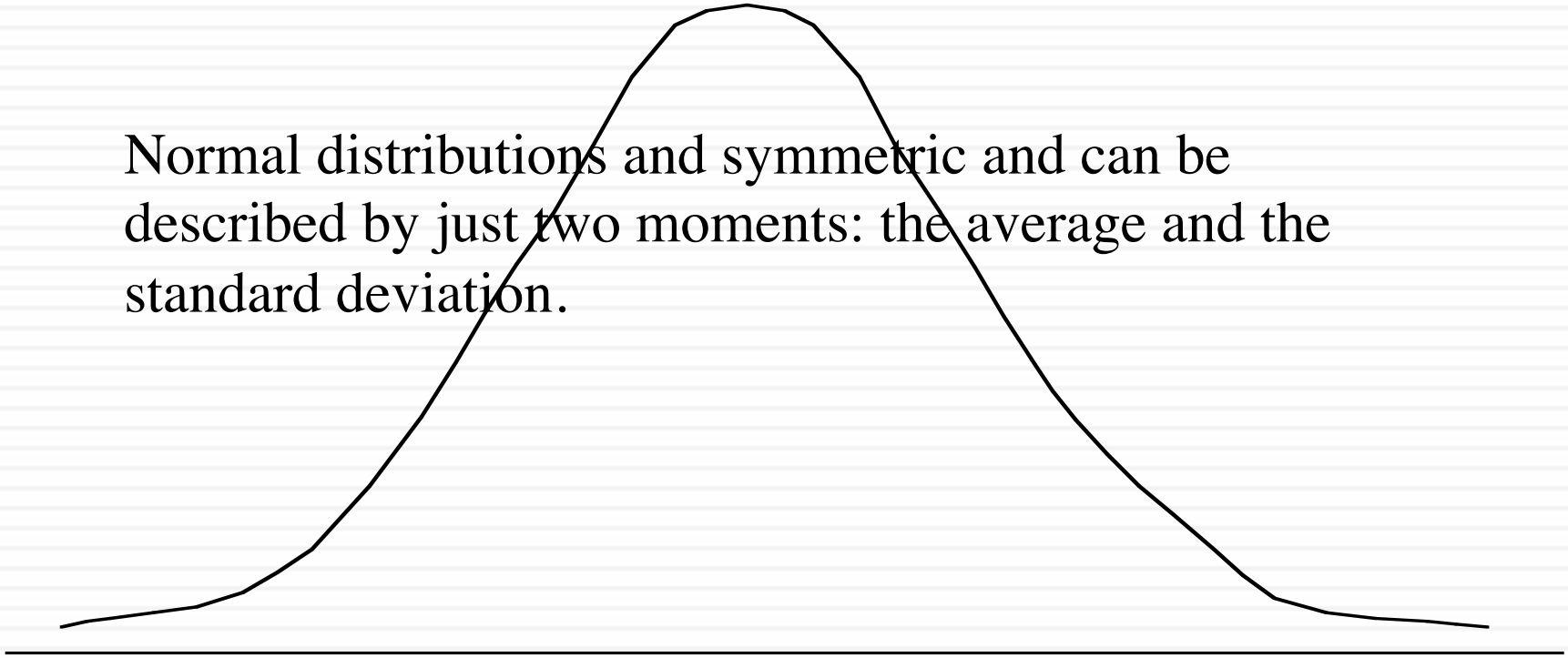
- When you are given lots of data, and especially when that data is contradictory and pulls in different directions, statistics help you make sense of the data and make judgments.

Summarizing Data

- As human beings, it is difficult for us to digest vast amounts of data. Data summaries help up by presenting the data in a more digestible form.
- One way to summarize data is visually, i.e., a distribution that reveals both what the observations share in common and where they are different.
- The other is with descriptive statistics: average, standard deviation etc..

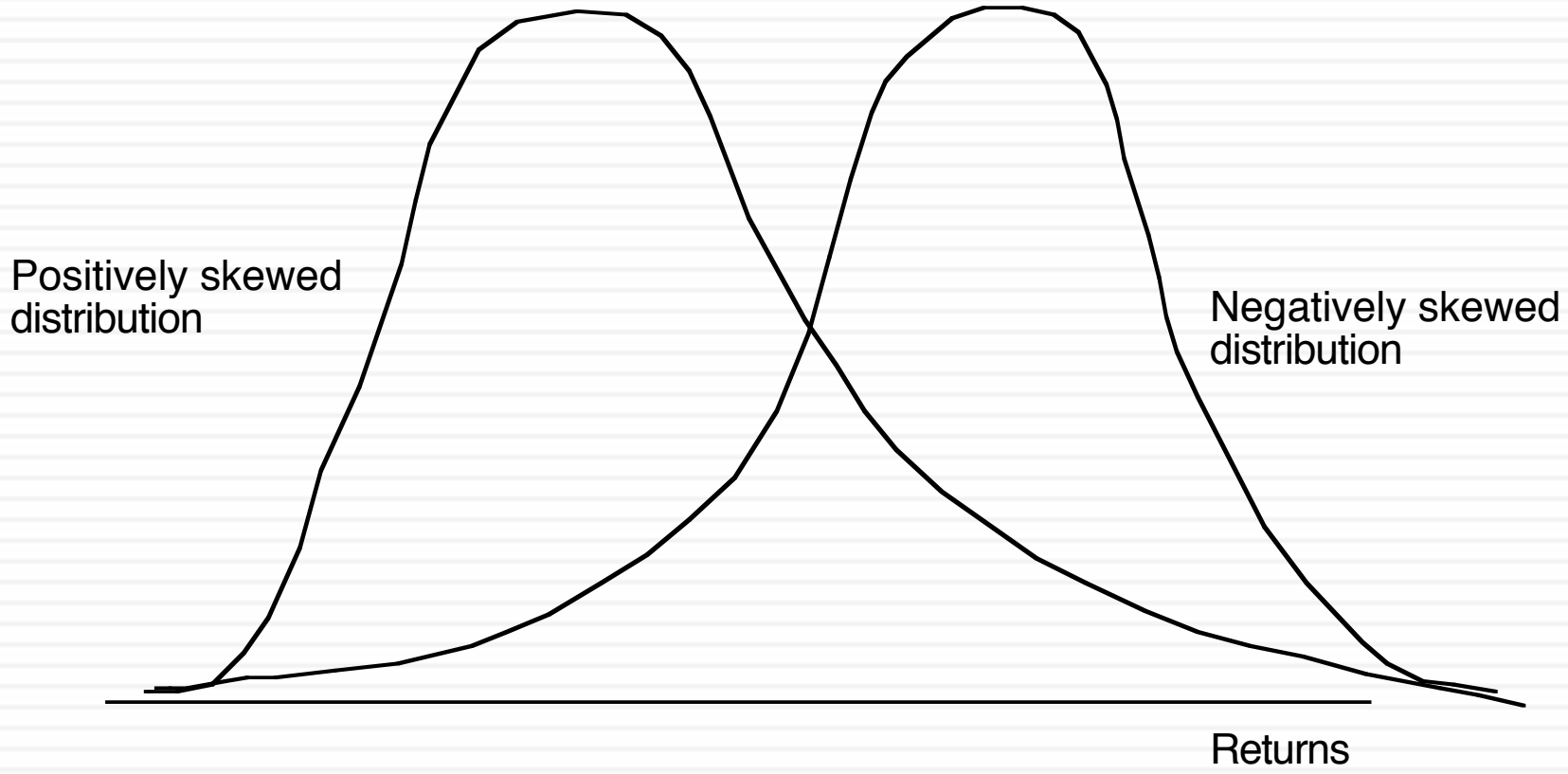
A Dream Distribution: The Normal

Figure A1.1: Normal Distribution



A more typical distribution: Skewed

Figure A1.2: Skewed Distributions



Summary Statistics: The most widely used!

For a data series, X_1, X_2, X_3, . . . X_n, where *n* is the number of observations in the series, the most widely used summary statistics are as follows:

1. The mean (m), which is the average of all of the observations in the data series.

$$Mean = \mu_X = \frac{\sum_{j=1}^{j=n} X_j}{n}$$

1. The variance, which is a measure of the spread in the distribution around the mean and is calculated by first summing up the squared deviations from the mean, and then dividing by either the number of observations (if it the population) or one less than that number (if it is a sample). The standard deviation is the square root of the variance.

$$\text{Variance} = \sigma_X^2 = \frac{\sum_{j=1}^{j=n} (X_j - \mu)^2}{n-1}$$

More summary statistics

- The median of a distribution is its exact midpoint, with half of all observations having values higher than that number and half lower. In a perfectly symmetric distribution (like the normal) the mean = median.
- If a distribution is not symmetric, the skewness (third moment) measures the direction (positive or negative) and degree of asymmetry.
- The kurtosis (fourth moment) measures the likelihood of extreme values in the data. A high kurtosis indicates that there are more observations that deviate a lot from the average.

Relationships between data: Covariance

- For two data series, X (X_1, X_2,) and Y(Y, Y. . .), the covariance provides a measure of the degree to which they move together and is estimated by taking the product of the deviations from the mean for each variable in each period.

$$\text{Covariance} = \sigma_{XY} = \frac{\sum_{j=1}^{j=n} (X_j - \mu_X)(Y_j - \mu_Y)}{n-1}$$

- The sign on the covariance indicates the type of relationship the two variables have. A positive sign indicates that they move together and a negative sign that they move in opposite directions.

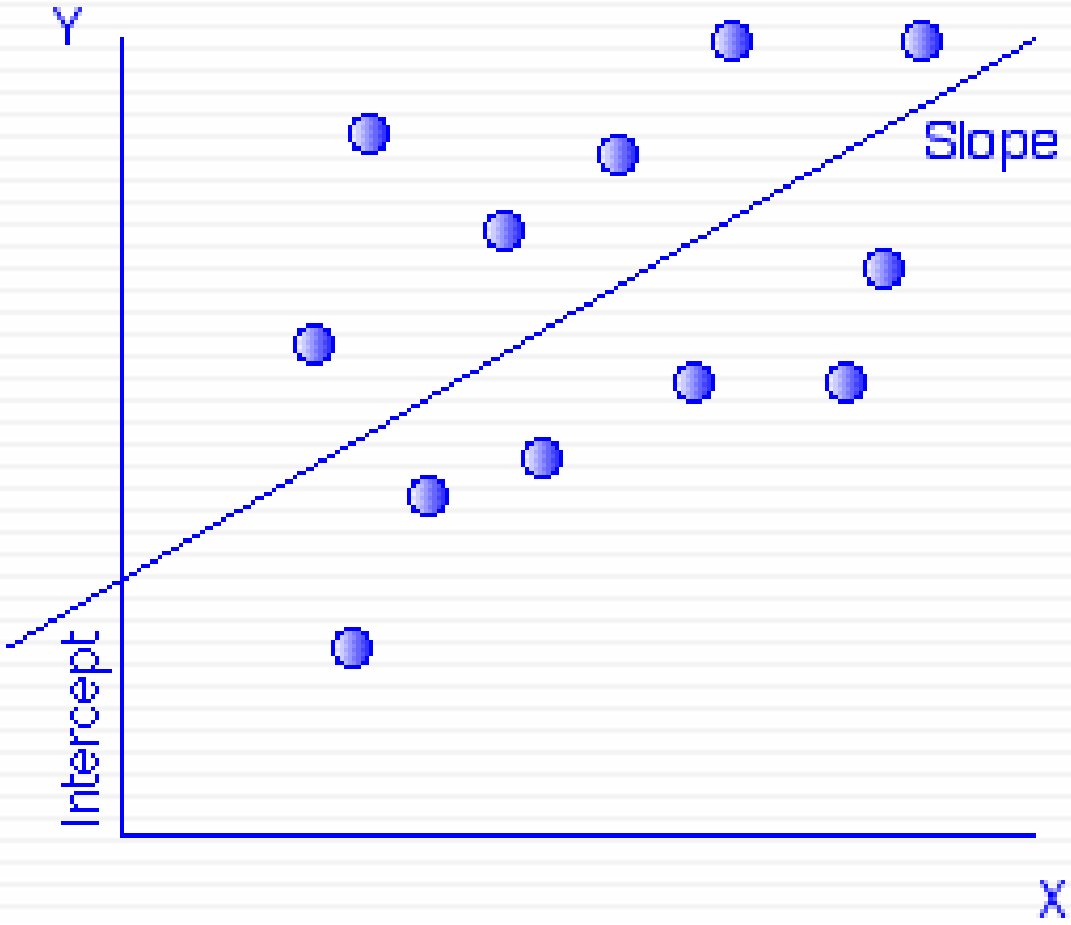
From covariance to correlation

- The correlation is the standardized measure of the relationship between two variables. It can be computed from the covariance.

$$\text{Correlation} = \rho_{XY} = \sigma_{XY} / \sigma_X \sigma_Y = \frac{\sum_{j=1}^{j=n} (X_j - \mu_X)\,(Y_j - \mu_Y)}{\sqrt{\sum_{j=1}^{j=n} (X_j - \mu_X)^2}\sqrt{\sum_{j=1}^{j=n} (Y_j - \mu_Y)^2}}$$

- A correlation close to zero indicates that the two variables are unrelated.
- A positive correlation indicates that the two variables move together, and the relationship is stronger as the correlation gets closer to one.
- A negative correlation indicates the two variables move in opposite directions, and that relationship gets stronger the as the correlation gets closer to negative one

Digging Deeper: Scatter Plots and Regressions



Reading a regression

- In a regression, we attempt to fit a straight line through the points that best fits the data. In its simplest form, this is accomplished by finding a line that minimizes the sum of the squared deviations of the points from the line.
- When such a line is fit, two parameters emerge—one is the point at which the line cuts through the *Y*-axis, called the intercept (a) of the regression, and the other is the slope (b) of the regression line:

$$Y = a + bX$$

- The slope of the regression measures both the direction and the magnitude of the relationship between the dependent variable (*Y*) and the independent variable (*X*). When the two variables are positively correlated, the slope will also be positive, whereas when the two variables are negatively correlated, the slope will be negative. The magnitude of the slope of the regression can be read as follows: For every unit increase in the dependent variable (*X*), the independent variable will change by *b* (slope).

How the intercept and slope are estimated

- The slope of the regression line is a logical extension of the covariance concept introduced in the last section. In fact, the slope is estimated using the covariance:

$$\text{Slope of the Regression} = b = \frac{\text{Covariance}_{YX}}{\text{Variance of X}} = \frac{\sigma_{YX}}{\sigma_X^2}$$

- The intercept (a) of the regression can be read in a number of ways. One interpretation is that it is the value that Y will have when X is zero. Another is more straightforward and is based on how it is calculated. It is the difference between the average value of Y, and the slope-adjusted value of X.

$$\text{Intercept of the Regression} = a = \mu_Y - b^* (\mu_X)$$

Measuring the noise in a regression

- The R^2 of the regression measures the proportion of the variability in the dependent variable (Y) that is explained by the independent variable (X). An R^2 value close to one indicates a strong relationship between the two variables, though the relationship may be either positive or negative.
- Another measure of noise in a regression is the standard error, which measures the "spread" around each of the two parameters estimated—the intercept and the slope.
- Dividing the coefficient (intercept or slope) by the standard error of the coefficient yields a t statistic which can be used to judge statistical significance.

Using Regressions for predictions

- The regression equation described in the last section can be used to estimate predicted values for the dependent variable, based on assumed or actual values for the independent variable. In other words, for any given *Y*, we can estimate what *X* should be:

$$X = a + b(Y)$$

- How good are these predictions? That will depend entirely on the strength of the relationship measured in the regression. When the independent variable explains a high proportion of the variation in the dependent variable (R^2 is high), the predictions will be precise. When the R^2 is low, the predictions will have a much wider range.

Simple to Multiple Regressions

- The regression that measures the relationship between two variables becomes a multiple regression when it is extended to include more than one independent variables (*X*1, *X*2, *X*3, *X*4 . . .)

$$Y = a + bX1 + cX2 + dX3 + eX4$$

- The R^2 still measures the strength of the relationship, but an additional R^2 statistic called the adjusted R^2 is computed to counter the bias that will induce the R^2 to keep increasing as more independent variables are added to the regression. If there are k independent variables in the regression, the adjusted R^2 is computed as follows:

$$\text{Adjusted R squared} = \frac{\left(\sum_{j=1}^{j=n}(Y_j - bX_j)^2\right)}{\text{n-k}}$$

Caveat Emptor on Regressions

- Both the simple and multiple regressions described in this section also assume linear relationships between the dependent and independent variables. If the relationship is not linear, we can either transform the data (either dependent on independent) to make the relationship more linear or run a non-linear regression.
- For the coefficients on the individual independent variables to make sense, the independent variable needs to be uncorrelated with each other, a condition that is often difficult to meet. When independent variables are correlated with each other, the statistical hazard that is created is called *multicollinearity*. In its presence, the coefficients on independent variables can take on unexpected signs (positive instead of negative, for instance) and unpredictable values.