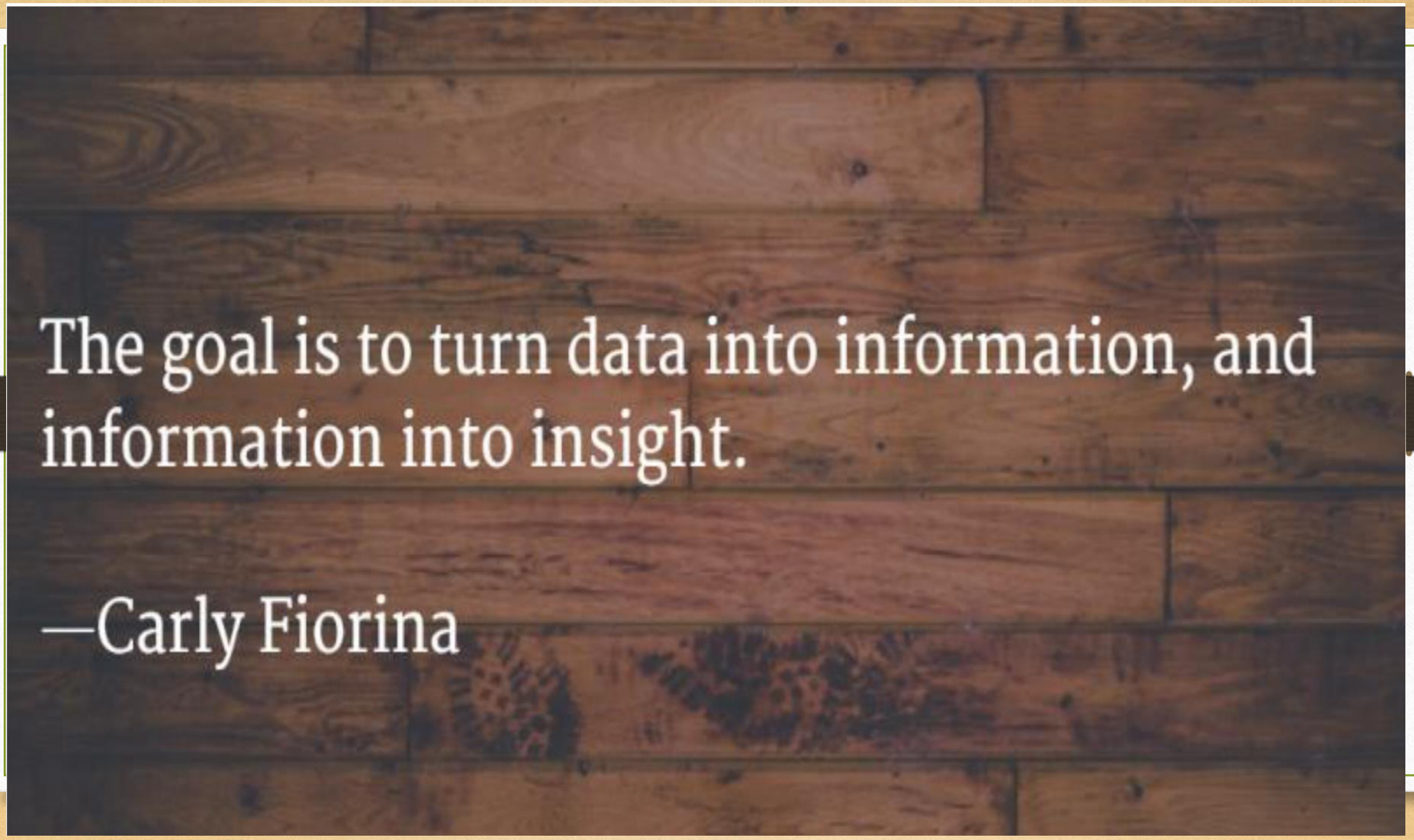
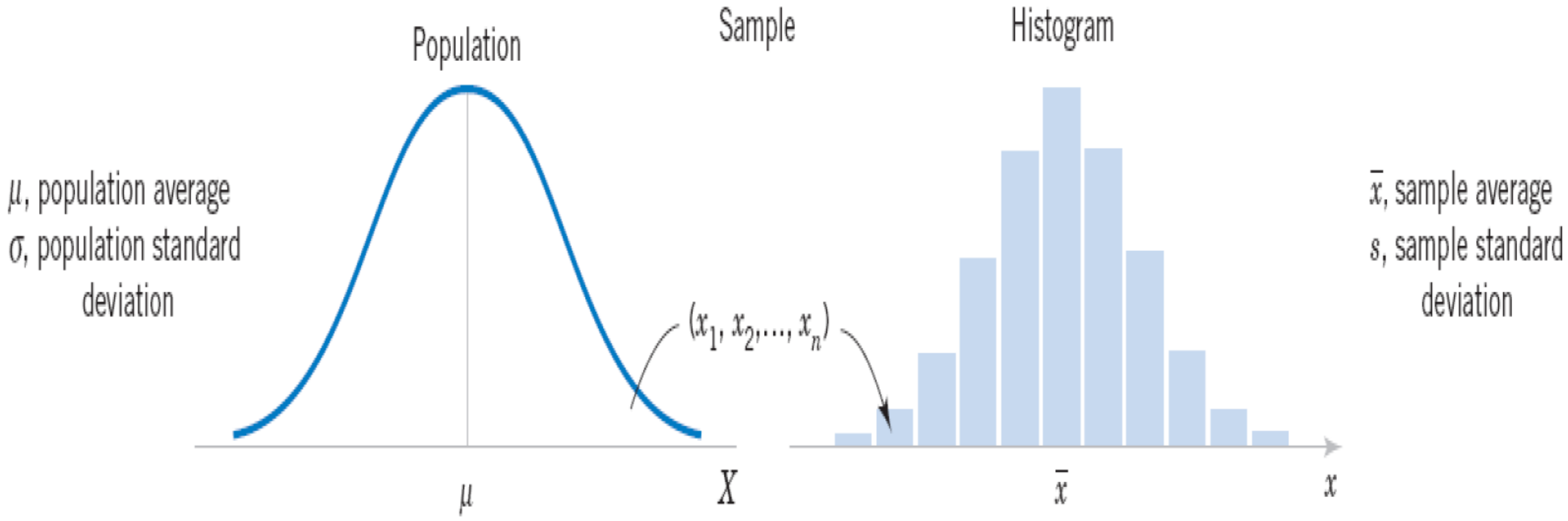
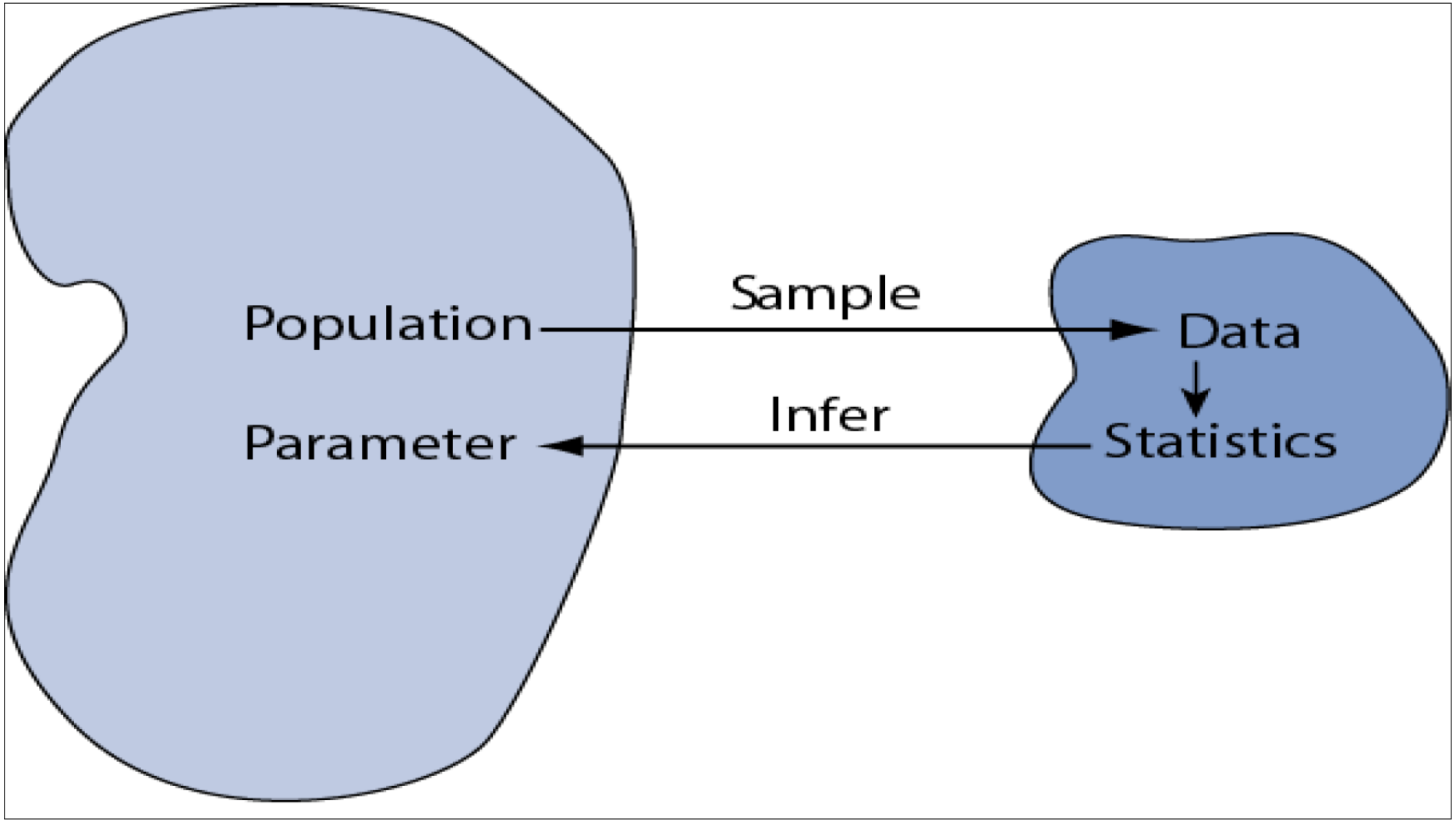
Linear Regression, Regression with Multiple Explanatory Variables


The goal is to turn data into information, and information into insight.
—Carly Fiorina

Relationship between a population and sample



Population and sample



Null and Alternative Hypotheses

Convert the research question to null and alternative hypotheses :

- The **null hypothesis (H_0)** is a claim of "no difference in the population"
- The **alternative hypothesis (H_a)** claims "H_0 is false"
- Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction)

General Procedure of Hypothesis Testing

1. **Parameter of interest:** From the problem context, identify the parameter of interest.
2. **Null hypothesis, H_0:** State the null hypothesis, H_0.
3. **Alternative hypothesis, H_1:** Specify an appropriate alternative hypothesis, H_1.
4. **Test statistic:** State an appropriate test statistic.
5. **Reject H_0 if:** Define the criteria that will lead to rejection of H_0.
6. **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
7. **Conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context. This could involve computing a P-value or comparing the test statistic to a set of critical values.

Sampling Distributions of a Mean

The **sampling distributions of a mean (SDM)** describes the behavior of a sampling mean

$$\bar{x} \sim N\left(\mu, SE_{\bar{x}}\right)$$

$$\text{where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

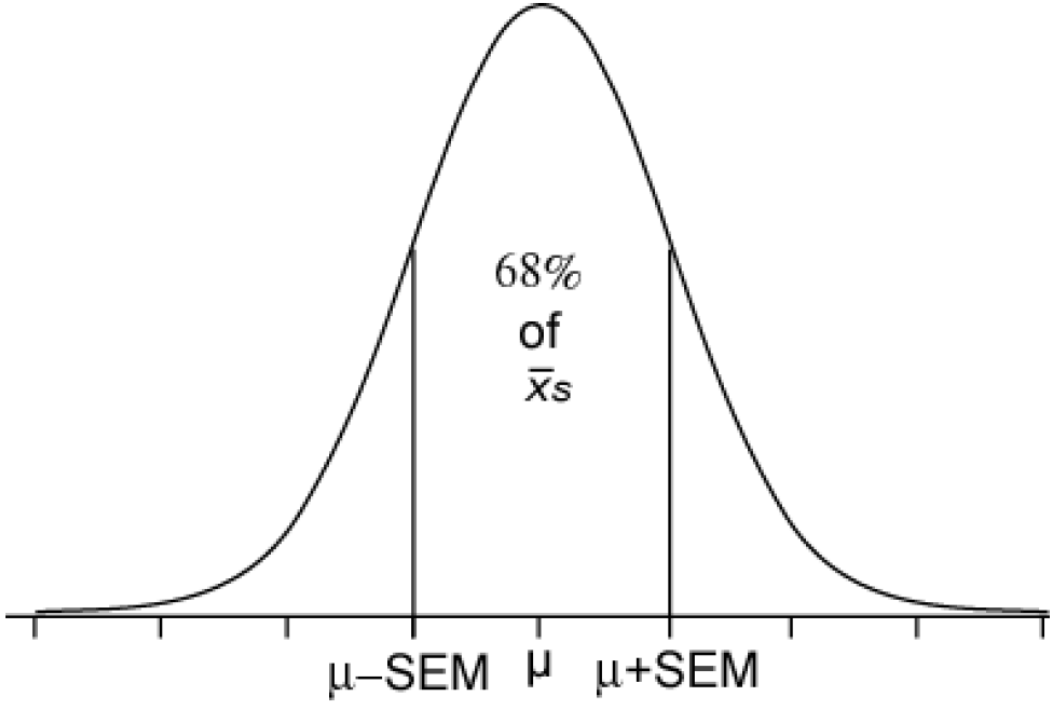


Fig. sdm&se68%.ai

Hypothesis Testing

- We like to think of statistical hypothesis testing as the data analysis stage of a **comparative experiment**, in which the Market Expert is interested, for example, in comparing the mean of a population to a specified value (e.g. mean pull strength).

- Suppose we are interested in knowing that price of a stock is more than 1500.

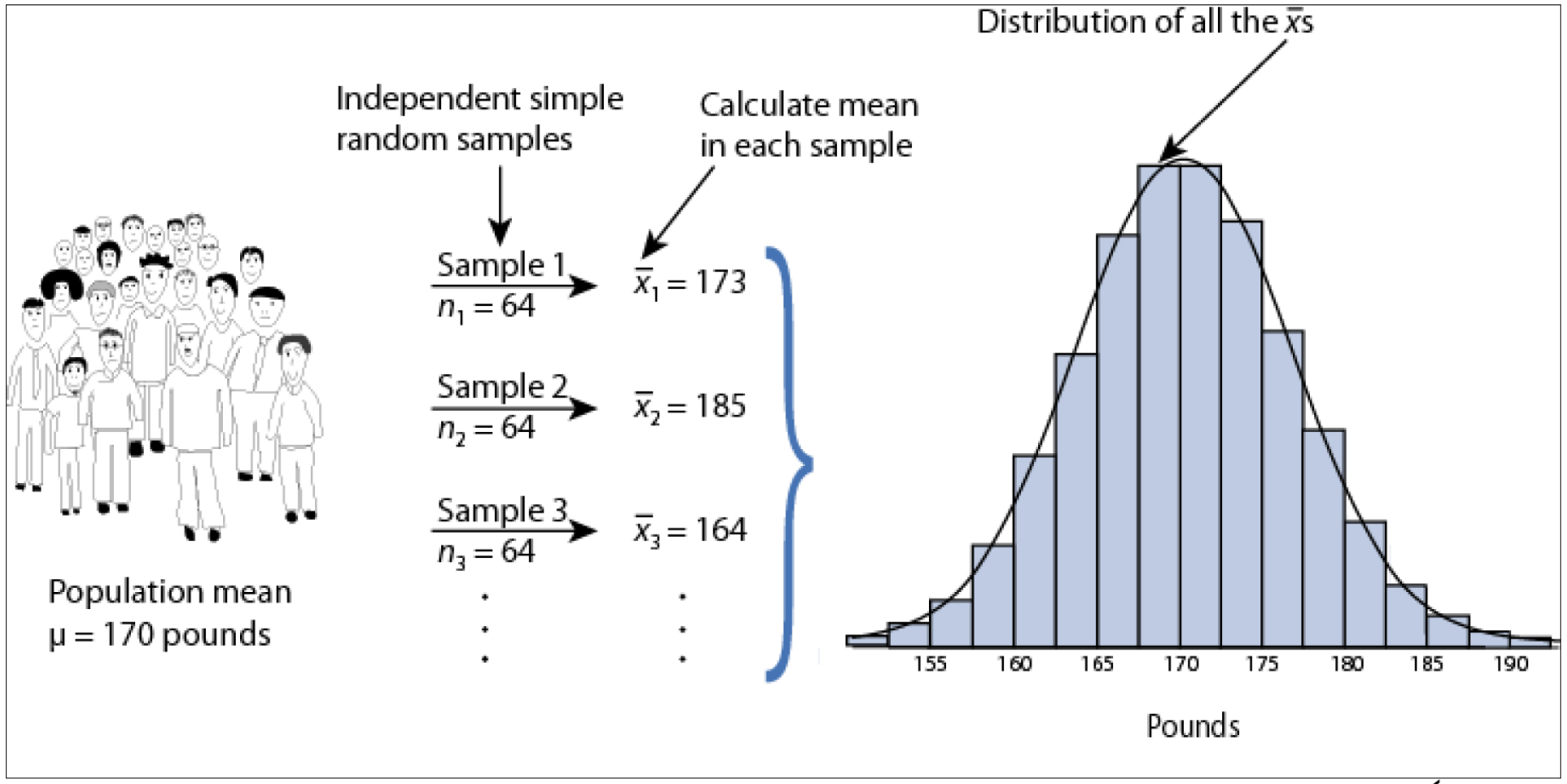
Hypothesis Testing

Two-sided Alternative Hypothesis

$H_0: \mu = 1500$

$H_0: \mu \neq 1500$

Reasoning Behind μz_{stat}

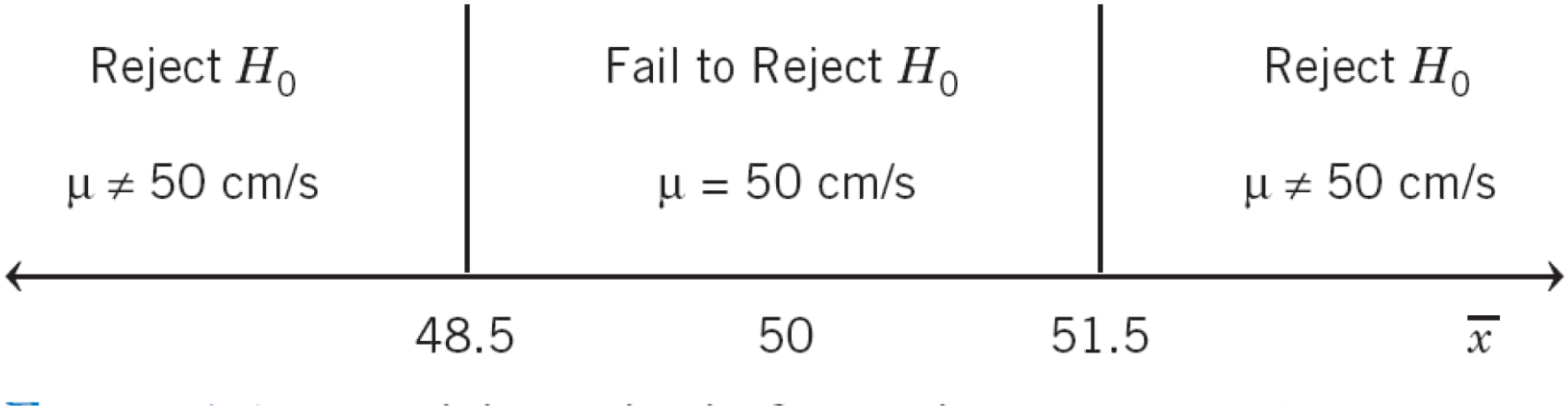


Sampling distribution of xbar under H_0: $\mu = 170$ for $n = 64 \Rightarrow$ $\bar{x} \sim N(170,5)$

Testing Statistical Hypotheses

$$H_0: \mu = 50 \text{ cm/s}$$

$$H_1: \mu \neq 50 \text{ cm/s}$$



Testing Statistical Hypotheses

Rejecting the null hypothesis H_0 when it is true is defined as a **type I error.**

Failing to reject the null hypothesis when it is false is defined as a **type II error.**

Testing Statistical Hypotheses

- The field of statistical inference consists of those methods used to make decisions or draw conclusions about a **population**.

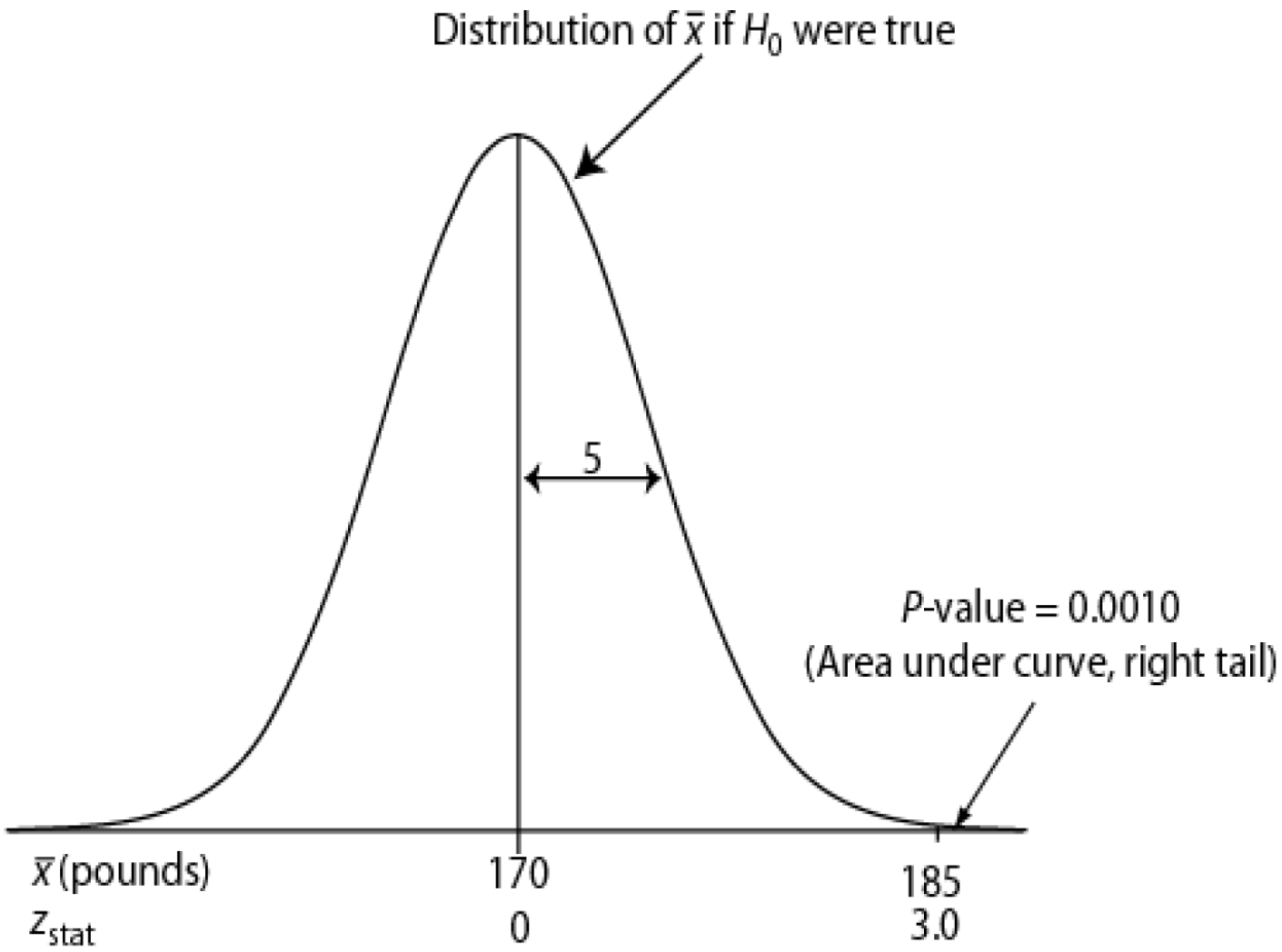
- These methods utilize the information contained in a **sample** from the population in drawing conclusions.

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

Sometimes the type I error probability is called the **significance level**, or the **α-error**, or the **size** of the test

P-value


Distribution of $\bar{x}$ if H_0 were true
5
P-value = 0.0010
(Area under curve, right tail)
$\bar{x}$ (pounds)
z_{stat}
170
0
185
3.0

P-Value

- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against H_0

- Small *P*-value $\Rightarrow$ strong evidence

One-Sample *z* Test

A. Hypothesis statements
   H_0: $\mu = \mu_0$ vs.
   H_a: $\mu \neq \mu_0$ (two-sided) or
   H_a: $\mu < \mu_0$ (left-sided) or
   H_a: $\mu > \mu_0$ (right-sided)

B. Test statistic

$$z_{stat} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \text{ where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

C. P-value: convert z_{stat} to P value

D. Significance statement (usually not necessary)

Question Normal Distribution

A companies share price is normally distributed with a mean weight of Rs 800 and a standard deviation of Rs 300 . A random sample of 16 days share price is taken. (a) What is the probability that the share price of the sample exceeds Rs 900 ?

μ= 800, $\bar{x}$ = 900 , S.E = 300/$\sqrt{16}$ =75

$$P(\bar{x} > 900) = \frac{900-800}{75} = 1.33$$

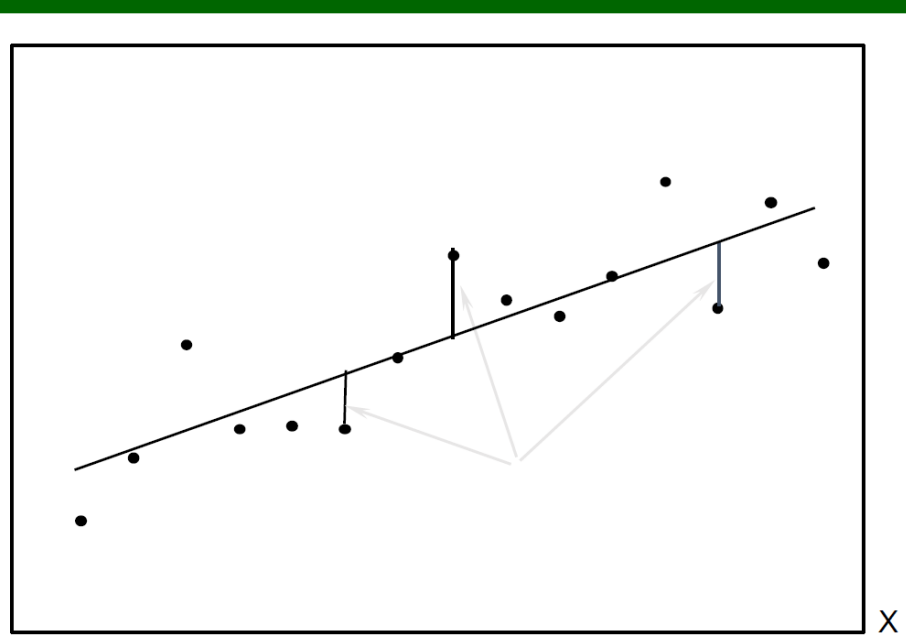
0.5-0.4082 = 0.0918 , Hence there is 9.18% probability.

Regression Analysis

- The statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called regression analysis.

Linear Regression model

Objective: The line that **BEST** fit the data.



Minimize the sum of difference between actual and fitted values?

Or

Minimize the sum of squares of difference between the actual and fitted value?

Regression Analysis

- **A time series is a set of observations *x* , *each one being recorded at a specific time t* .**
- Time series data, as the name suggests, are data that have been collected over a period of time on one or more variables.

- **Problems that can solved with Time-Series:**
  - How the value of a country's stock index has varied with that country's macroeconomic fundamentals.
  - How the value of a company's stock price has varied when it announced the value of its dividend payment.
  - The effect on a country's exchange rate with increase in its trade deficit.
  - How interest rate are determined.
  - Finding out risk in an asset class.

Regression Analysis

You are fitting a below mentioned straight -line equation.

$$y_i = \beta_1 + \beta_2 x_2$$

where ß1 and ß2 are unknown but fixed parameters known as the **regression coefficients.**

In regression analysis our interest is in estimating the PRFs**.**

Regression Analysis

THE COEFFICIENT OF DETERMINATION- R2

A MEASURE OF "GOODNESS OF FIT"

We now consider the **goodness of fit** of the fitted regression line to a set of data; that is, we shall find out how "well" the sample regression line fits the data.

The **coefficient of determination** *r* 2 (two-variable case) or *R*2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data.

$$R^2 = \frac{ESS}{TSS}$$

Where ESS = Explained sum of square
TSS = Total sum of square

Regression Analysis

Output Interpretation

Regression Statistics										
Multiple R	0.983559									
R Square	0.967389									
Adjusted R Square	0.964672									
Standard Error	19.00868									
Observations	14									
ANOVA										
	df	SS	MS	F	Significance F					
Regression	1	128625.3	128625.3	355.9772	2.75E-10					
Residual	12	4335.96	361.33							
Total	13	132961.2								
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%		
Intercept	-22.5464	10.43676	-2.16029	0.051685	-45.2861	0.193368	-45.2861	0.193368		
X Variable 1	3.269721	0.1733	18.86736	2.75E-10	2.892132	3.64731	2.892132	3.64731		

Regression Analysis

Null Hypothesis of Regression Co-efficient H_0 = All coefficients =0

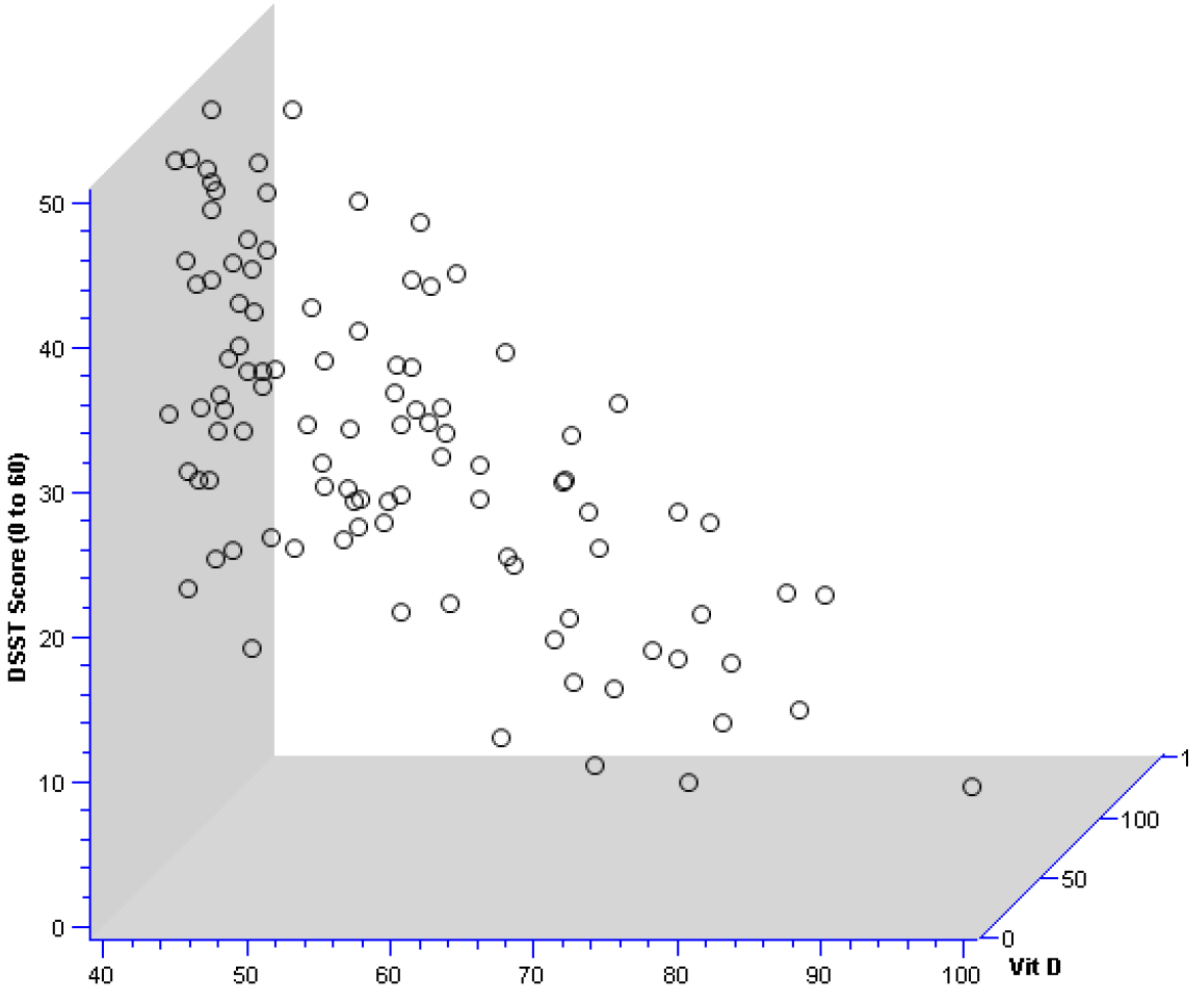
Significance testing…

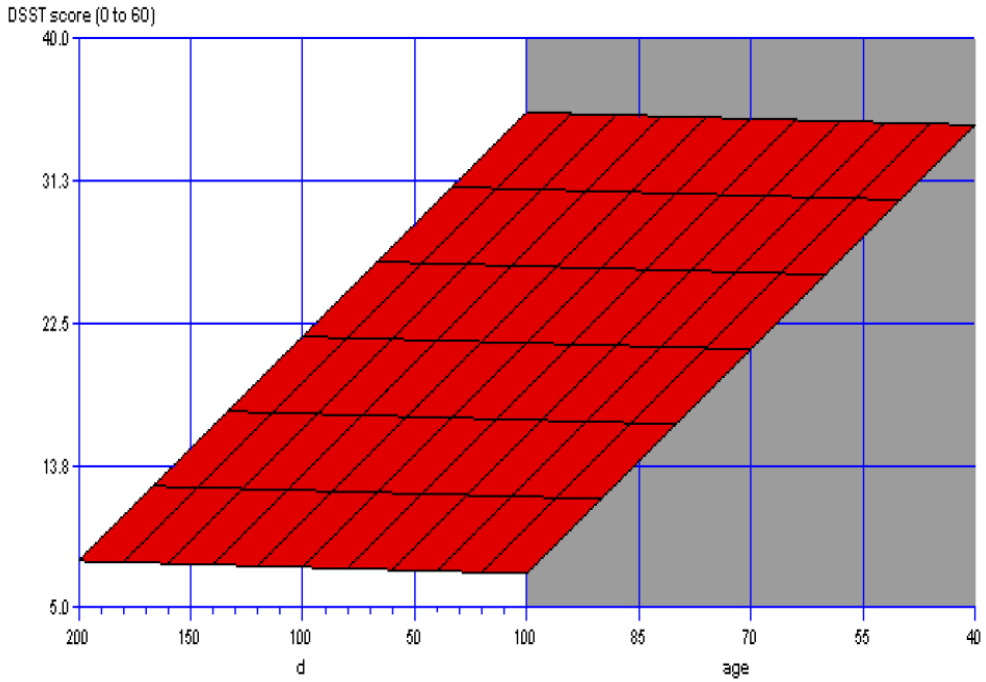
H0: $\beta 1 = 0$ (no linear relationship)

H1: $\beta 1 \neq 0$ (linear relationship does exist)

$$T_{n-2} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

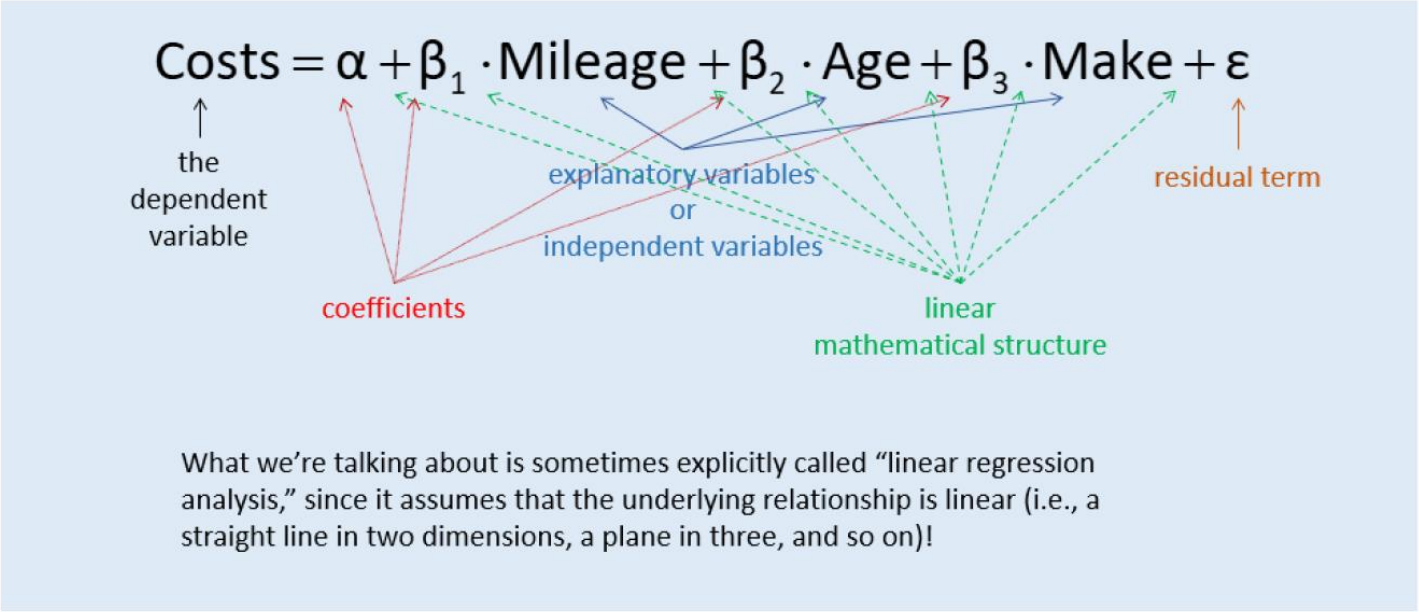
Multiple linear regression (We fit a Plane)





The Regression Model

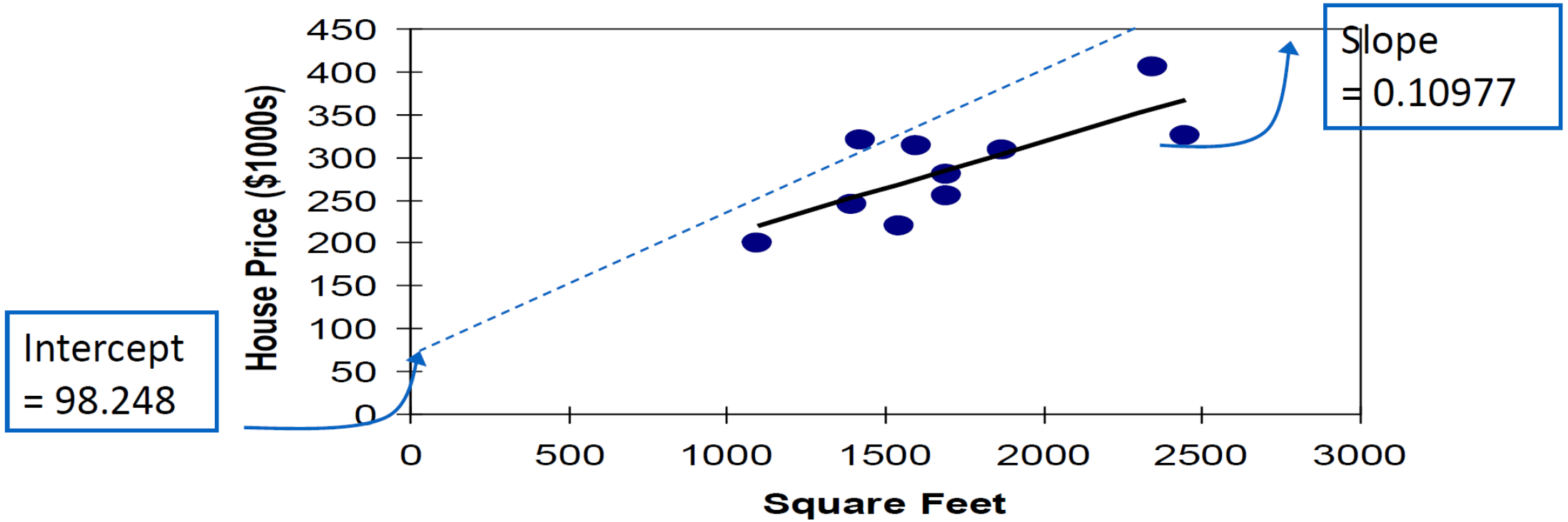
$$\text{Costs} = \alpha + \beta_1 \cdot \text{Mileage} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Make} + \varepsilon$$



What we're talking about is sometimes explicitly called "linear regression analysis," since it assumes that the underlying relationship is linear (i.e., a straight line in two dimensions, a plane in three, and so on)!

The Regression Model

- House price model: scatter plot and regression line



Interpretation of the Intercept, b_0

$$\text{Price} = 98.24833 + 0.10977\ (\text{sales})$$

b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

विद्या परं देवतम्
IIM
VISAKHAPATNAM

Regression Estimator

Regression estimator should be BLUE

B= Best
L= Linear
U= Unbiased
E= Estimator