SEC “ 5 ”

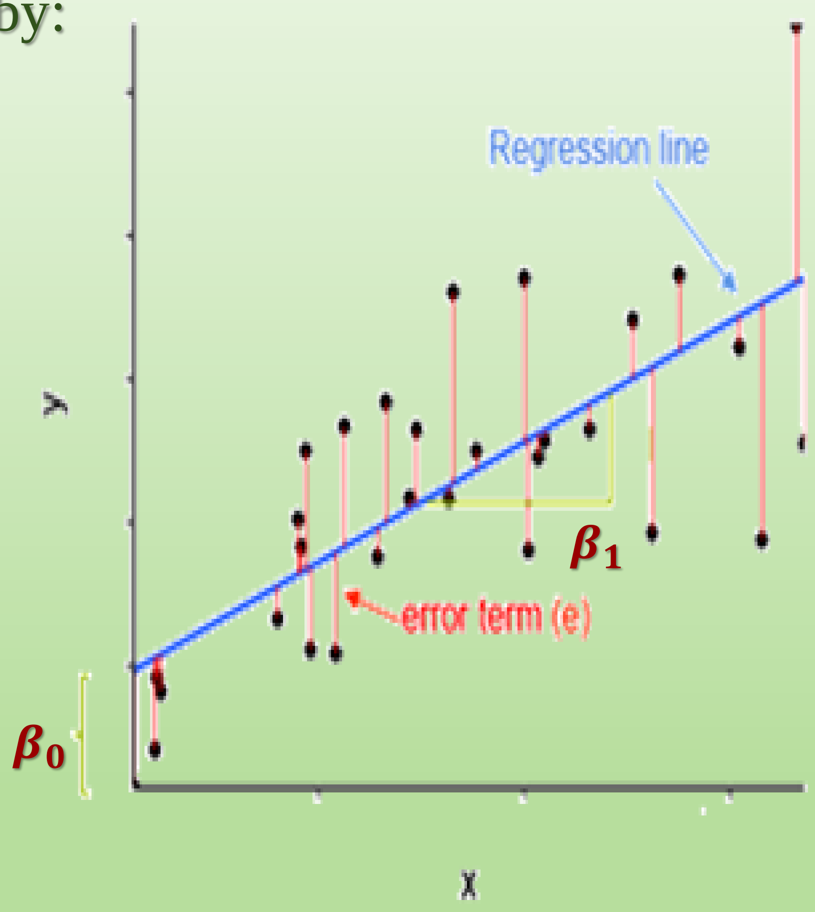
SIMPLE LINEAR REGRESSION

BAYES' THEOREM

Simple Linear Regression

- **Linear regression models** are used to describe or predict the relationship between two variables " x and y ". The simple linear regression model is represented by:

$$y = \beta_0 + \beta_1 x + e$$



y : The factor that is being predicted (the factor that the equation solves for) is called the dependent variable.

x : The factors that are used to predict the value of the dependent variable are called the independent variables.

e : Is the error of the estimate. The error term is used to account for the variability in y that cannot be explained by the linear relationship between x and y.

β_0 : Is the y-intercept of the regression line.

β_1 : Is the slope.

The Estimated Linear Regression Equation

- **In practice,** the parameter of the population values generally are not known so they must be estimated by using data from a sample of the population. The population parameters are estimated by using sample statistics. The sample statistics are represented by b_0 and b_1. When the sample statistics are substituted for the population parameters, the estimated regression equation is formed as follow:

$$E(y) = \hat{y} = b_0 + b_1 x, \qquad where$$

Sample mean of y

Sample mean of x

$$b_0 = E(\beta_0) = \bar{y} - b_1\bar{x},$$

and

Correlation coefficient

Standard deviation of y

$$b_1 = E(\beta_1) = r\left(\frac{s_y}{s_x}\right)$$

Standard deviation of x

- **the error in the predicted value of y at a certain value of x:**

  Error = $|\hat{y} - y|$

- **The coefficient of determination r^2:**

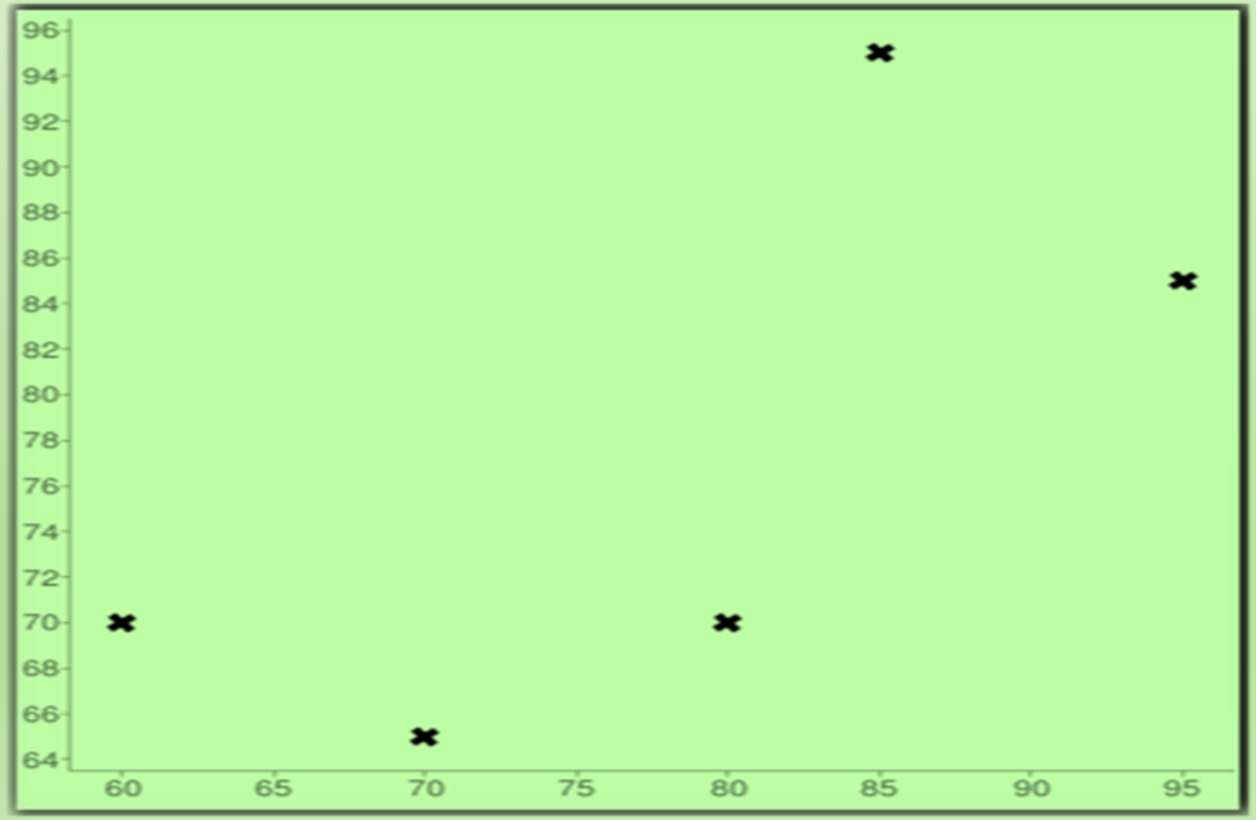
  how well does the regression equation fit the data. This means that % of the variation in y can be described by x.

Sheet (3)

12. Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- Draw the scatter plot representing the data

Student	x_i	y_i
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70



- What linear regression equation best predicts statistics performance, based on math aptitude scores?
  - ❑ $n = 5$
  - ❑ $\overline{X} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{95+\cdots+60}{5} = 78$
  - ❑ $s_x = \sqrt{\frac{\sum_{i=1}^{5}(x-\overline{x})^2}{n-1}} = \sqrt{\frac{(95-78)^2+\cdots+(95-78)^2}{4}} = 13.5093$
  - ❑ $\overline{y} = \frac{\sum_{i=1}^{5} y_i}{n} = \frac{85+\cdots+70}{5} = 77$
  - ❑ $s_y = \sqrt{\frac{\sum_{i=1}^{5}(y-\overline{y})^2}{n-1}} = \sqrt{\frac{(85-77)^2+\cdots+(70-77)^2}{4}} = 12.5499$

- $\hat{y} = b_0 + b_1 x$

- $b_1 = r\left(\frac{s_y}{s_x}\right)$

$$= 0.6931\left(\frac{12.5499}{13.5093}\right)$$

$$= 0.644$$

- $b_0 = \bar{y} - b_1\bar{x} = 77 - 0.644 * 78$

$$= 26.768$$

$$\hat{y} = 26.768 + 0.644\,x$$

x	$z_x = \frac{x-78}{13.5093}$	y	$z_y = \frac{y-77}{12.5499}$	$z_x z_y$
95	1.2584	85	0.6375	0.8022
85	0.5182	95	1.4343	0.7433
80	0.1480	70	−0.5578	−0.0826
70	−0.5922	65	−0.9562	0.5663
60	−1.3324	70	−0.5578	0.7432
			Total =	2.7724

$$r = \frac{\sum z_x z_y}{n-1} = \frac{2.7724}{4} = 0.6931$$

- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

At $x = 80$ $\longrightarrow$ $\hat{y}_{80} = 26.768 + 0.644 * 80 = 78.288.$

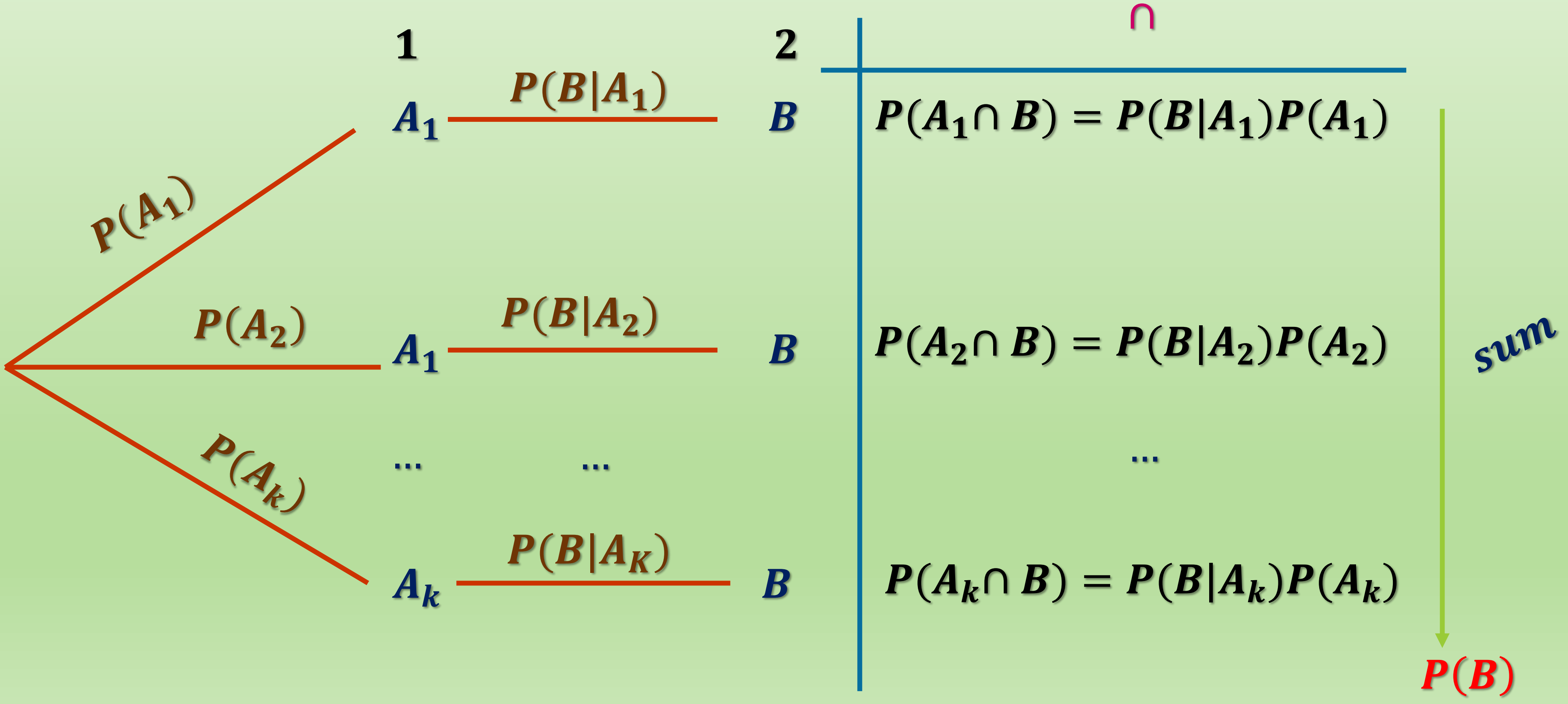
$Error = |\hat{y} - y| = |78.288 - 70| = 8.288$

- How well does the regression equation fit the data? (hint: use the coefficient of determination to answer this question).

$r^2 = (0.6931)^2 = 0.4804 \times 100 = 48.04\%$ **of the variation in y can be described by x.**

Bayes' Rule

- Bayes' theorem, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence



- $P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_k \cap B)$

  $= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)$

  $= \sum_{j=1}^{k} P(B|A_j)P(A_j)$ “ Total Probability ” .

- $P(A_i|B) = \dfrac{P(A_i \cap B)}{P(B)} = \dfrac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}$, $i = 1, 2, \ldots, k$

Sheet (3) [Revision on Probability]

4. All tractors made by a company are produced on one of three assembly lines, named Red, White, and Blue. The chances that a tractor will not start when it rolls off of a line are 6%, 11%, and 8% for lines Red, White, and Blue, respectively. 48% of the company's tractors are made on the Red line and 31% are made on the Blue line.

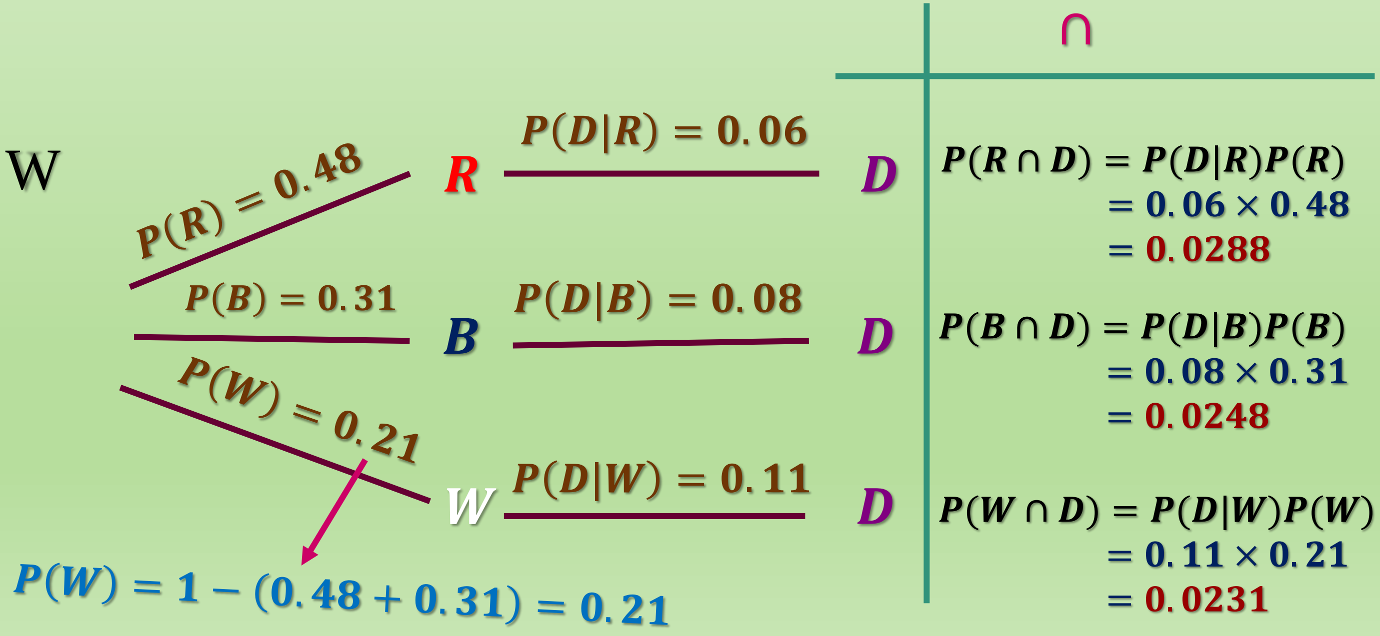
(a) What fraction of the company's tractors do not start when they roll off of an assembly line?

Let:

Red Line: R, Blue Line: B, White Line: W

Not Start " Defective ": D

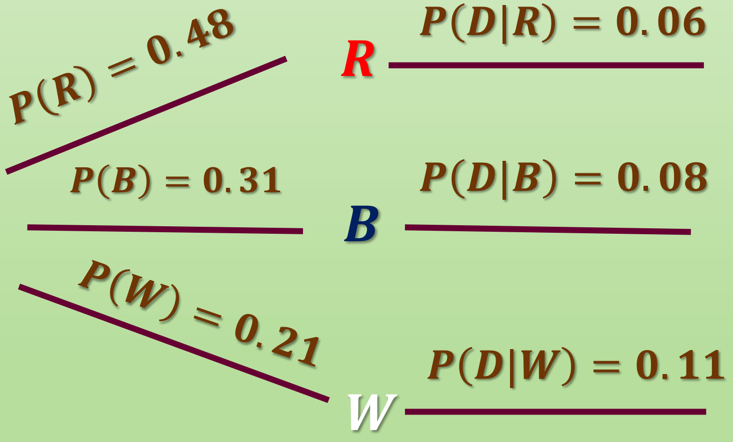
$$P(D)$$
$$= P(R \cap D) + P(B \cap D) + P(W \cap D)$$
$$= 0.0288 + 0.0248 + 0.0231$$
$$= 0.0767 \quad \times 100$$
$$= 7.67 \approx 8\%$$



(b) What is the probability that <u>a tractor came from the red company given that it was defective</u>?

$$P(R|D) = \frac{P(R \cap D)}{P(D)}$$

$$= \frac{0.0288}{0.0767} = 0.3755$$



$\cap$

$$P(R \cap D) = P(D|R)P(R) = 0.06 \times 0.48 = 0.0288$$

$$P(B \cap D) = P(D|B)P(B) = 0.08 \times 0.31 = 0.0248$$

$$P(W \cap D) = P(D|W)P(W) = 0.11 \times 0.21 = 0.0231$$

$+$

$$P(D) = 0.0767$$

Sheet (3) [Revision on Probability]

2. **A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?**

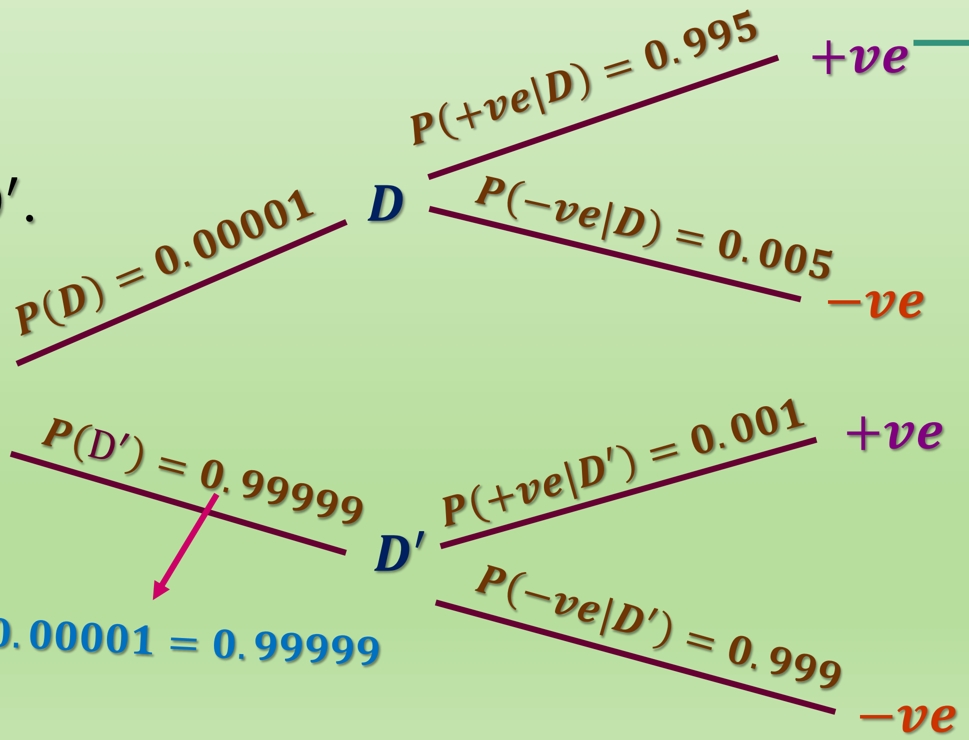
Let:

The person with the disease: D.

The person without the disease: D′.

Note that:

$P(A|E) + P(B|E) + \cdots + P(Z|E) = 1$



$\boldsymbol{P}(\mathrm{D}') = 1 - 0.00001 = 0.99999$

$\cap$

$$\begin{aligned} P(D \cap +ve) &= P(+ve|D)P(D) \\ &= 0.995 \times 0.00001 \\ &= 9.95 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} P(D \cap -ve) &= P(-ve|D)P(D) \\ &= 0.005 \times 0.00001 \\ &= 5 \times 10^{-3} \end{aligned}$$

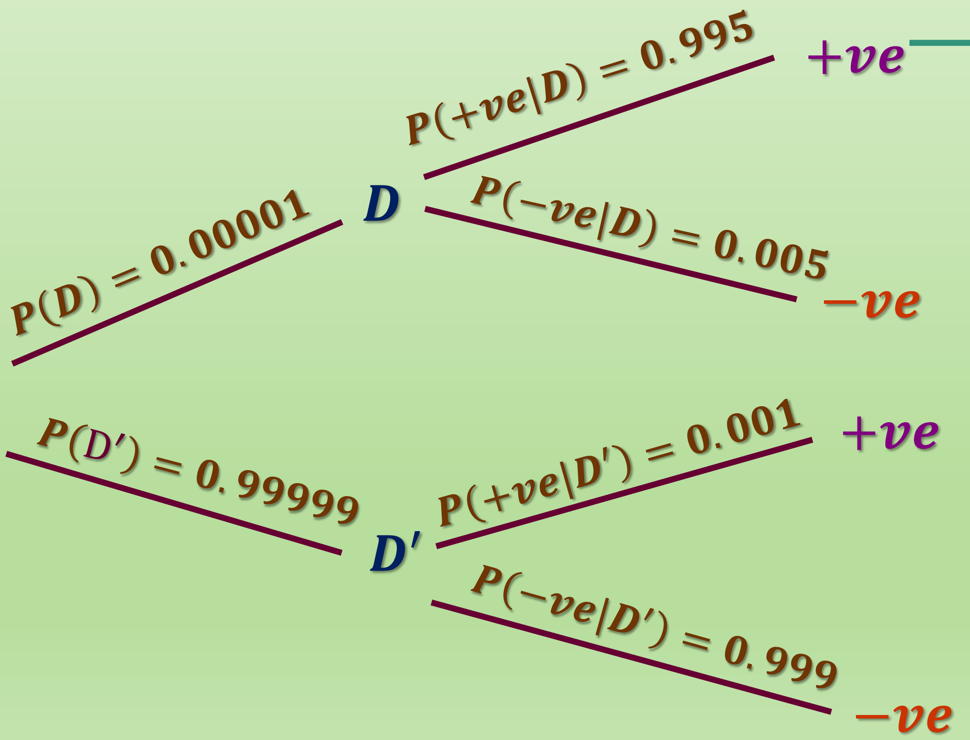
$$\begin{aligned} P(D' \cap +ve) &= P(+ve|D')P(D') \\ &= 0.001 \times 0.99999 \\ &= 9.9999 \times 10^{-4} \end{aligned}$$

$$\begin{aligned} P(D' \cap -ve) &= P(-ve|D')P(D') \\ &= 0.999 \times 99999 \\ &= 0.99899001 \end{aligned}$$

what is the likelihood that an individual who tests positive will actually have the disease?

$$P(D|+ve) = \frac{P(D \cap +ve)}{P(+ve)} = \frac{P(D \cap +ve)}{P(D \cap +ve) + P(D' \cap +ve)}$$

$$= \frac{9.95\times10^{-6}}{9.95\times10^{-6} + 9.9999\times10^{-4}} = 9.8521 \times 10^{-3} = 0.0098521$$



$\cap$

$$P(D \cap +ve) = P(+ve|D)P(D) = 0.995 \times 0.00001 = 9.95 \times 10^{-6}$$

$$P(D \cap -ve) = P(-ve|D)P(D) = 0.005 \times 0.00001 = 5 \times 10^{-3}$$

$$P(D' \cap +ve) = P(+ve|D')P(D') = 0.001 \times 0.99999 = 9.9999 \times 10^{-4}$$

$$P(D' \cap -ve) = P(-ve|D')P(D') = 0.999 \times 99999 = 0.99899001$$