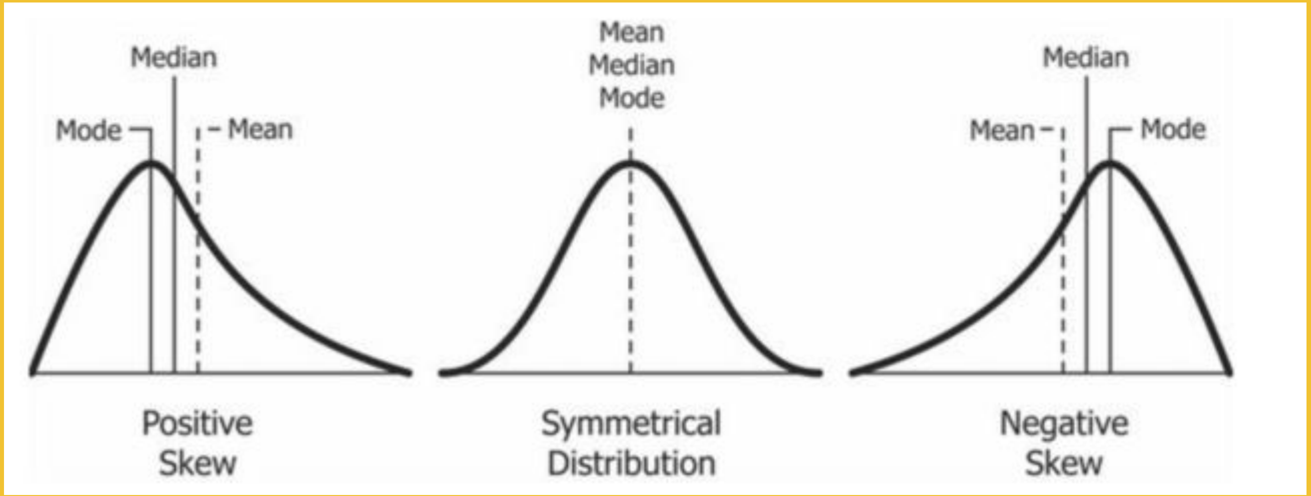


Econ 2250: Stats for Econ

Fall 2022

[Source for pic stats above.]

- # Review
- **Summary Statistics**
  - **Sum**
  - **Average**
  - **Median**
    - **Even and odd**
  - **Deviation**
  - **Variance**
  - **Standard Deviation**
- **Probability**
  - **Joint**
  - **Conditional**
- **Modeling**
  - **Covariance**
    - **Correlation**

Summary Statistics

	total_rooms	total_bedrooms	median_house_value
count	17000.000000	17000.000000	17000.000000
mean	2643.664412	539.410824	207300.912353
std	2179.947071	421.499452	115983.764387
min	2.000000	1.000000	14999.000000
25%	1462.000000	297.000000	119400.000000
50%	2127.000000	434.000000	180400.000000
75%	3151.250000	648.250000	265000.000000
max	37937.000000	6445.000000	500001.000000

Remember that median is middle

$$\text{if } n \text{ is odd, } \mathrm{median}(x) = x_{(n+1)/2}$$

$$\text{if } n \text{ is even, } \mathrm{median}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

Summary Operator

$$\sum X = x_1 + x_2 + \ldots + x_n$$

Summary

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n$$

Summary Operator Properties

1.) $\sum_{i=1}^{n} c = nc$

2.) $\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$

3.) For any constant a and b: $\sum_{i=1}^{n} (ax_i + by_i) = a \sum_{i=1}^{n} x_i + b \sum_{j=1}^{n} y_i$

Gotchas! Be Careful

$$\sum_{i}^{n} \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}$$

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2$$

Summary Operation (SIGMA?)

- And we can divide

```
x = [3,12,4]
y = [2,9,1]
sum(x*y)/sum(x**2)
```

$$\frac{\sum x_i y_i}{\sum x_i^2} = \frac{x_1 * y_1 + x_2 * y_2 + x_3 * y_3}{x_1^2 + x_2^2 + x_3^2}$$

$$= \frac{3 * 2 + 12 * 9 + 4 * 1}{3^2 + 12^2 + 4^2}$$

$$= \frac{6 + 108 + 4}{9 + 144 + 16}$$

$$= \frac{118}{169} = 0.6982249$$

Summary Operator Property 3:

For any constant a and b: $\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{j=1}^{n} y_i$

$$\sum_{i=1}^{n}(ax_i + by_i) =$$

$$(ax_1 + by_1) + (ax_2 + by_2) + (ax_3 + by_3) =$$

$$ax_1 + by_1 + ax_2 + by_2 + ax_3 + by_3 =$$

$$(ax_1 + by_1) + (ax_2 + by_2) + (ax_3 + by_3) =$$

$$a(x_1 + x_2 + x_3) + b(y_1 + y_2 + y_3) =$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

Show that…using x = (1,2,3)

$$\sum_{i}^{n} \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}$$

And

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

Expected Value Operator

$$E(x) = \Sigma\, x_i * Pr(x_i)$$

Expected Value Operator Property 2:

$$E(aX + b) = E(aX) + E(b) = aE(X) + b.$$

x = [3,6,2]

p(x) = ⅓

a = 5

b = 4

E(aX + b) = E(aX) + E(b)

$= ax_1*p(x_1) + ax_2*p(x_2) + ax_3*p(x_3) + b = a(x_1*p(x_1) + x_2*p(x_2) + x_3*p(x_3)) + b$

= a(E(x)) + b = 5(3*⅓ + 6*⅓ + 2*⅓) + 4 = 22⅓

Variance

$$V(X) = E((X - E(X))^2)$$

From the example above

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu_x)^2$$

x	mu	demean	demean_sq
3	6.3	-3.3	11.1
12	6.3	5.7	32.1
4	6.3	-2.3	5.4

Sum of demean_sq = 11.1 + 32.1 + 5.4 = 48.6

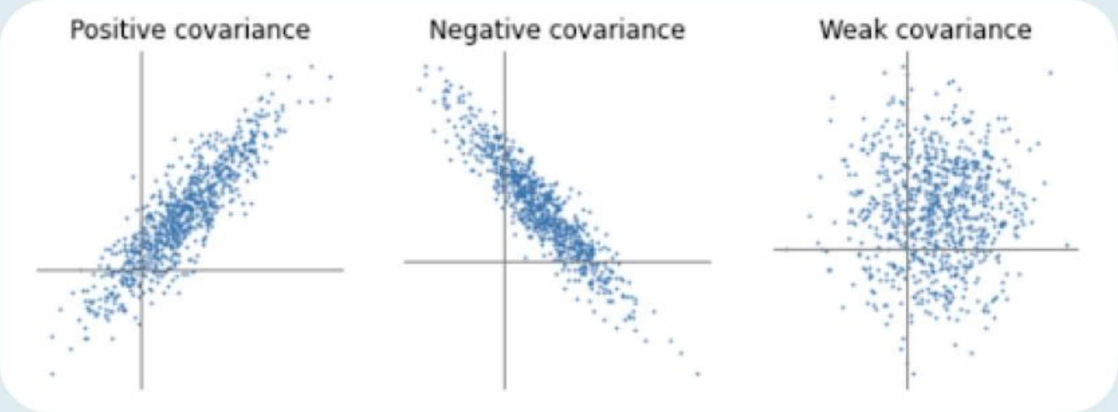
48.6/3 = 16.2

But, notice our deviations (-3.3, 5.7, -2.3), 16.2 is an awful absolute value estimate. That is because we squared the errors, and x is in levels (not squared).

Standard Deviation = square root of σ^2, sqrt(16.2) = 4.02

Nice correlation app

https://shiny.rit.albany.edu/stat/rectangles/



Covariance

$Cov(X, Y)$ = E[(X-E(X)(Y-E(Y)]

image

Covariance

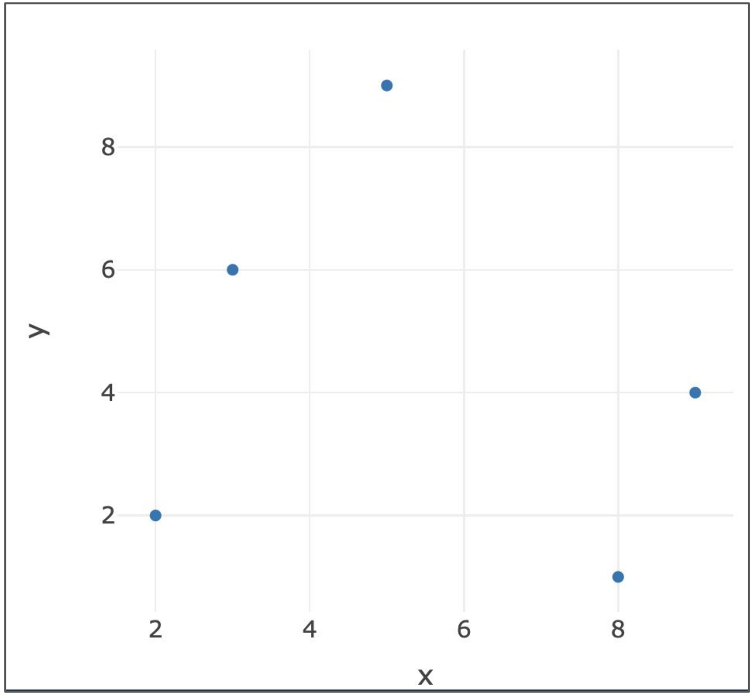
$$Cov(x, y) = E[(X - E(X))(Y - E(Y))]$$
$$= \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n - 1}$$

Example covariance

$$\frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n - 1}$$

x	y	demean_x	demean_y	demean_x*demean_y	
3	6	-2.4	1.6	-3.84	
5	9	-0.4	4.6	-1.84	
2	2	-3.4	-2.4	8.16	
8	1	2.6	-3.4	-8.84	
9	4	3.6	-0.4	-1.44	
				-7.8	sum
				-1.95	sum/(n-1)

mean_y	4.4
mean_x	5.4



Correlation

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Correlation

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \quad \text{if } \sigma_X \sigma_Y > 0$$

$$r_{xy} \stackrel{\text{def}}{=} \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2 \sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}$$

Example

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x	y	demean_x	demean_x_sq	demean_y	demean_y_sq	demean_x*demean_y	
3	6	-2.4	5.76	1.6	2.56	-3.84	
5	9	-0.4	0.16	4.6	21.16	-1.84	
2	2	-3.4	11.56	-2.4	5.76	8.16	
8	1	2.6	6.76	-3.4	11.56	-8.84	
9	4	3.6	12.96	-0.4	0.16	-1.44	
			37.2		**41.2**	-7.8	sum
						-1.95	sum/(n-1)

mean_y	4.4
mean_x	5.4

numerator	**-1.95**	-1.95	**-0.22**	**correlation**
denom	sqrt(**37.2** + **41.2**)	8.85		

Probability

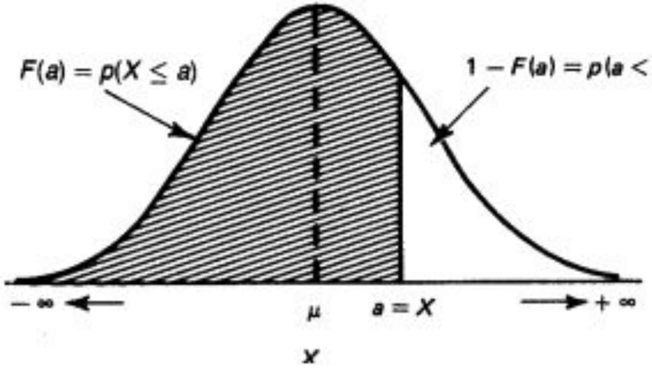
Probability notation

- Capital letters, such as X, Y, and Z, are used to denote random variables.
- Lowercase letters, such as x, y, z and a, b, c are used to denote particular values that the random variable can take on.
- Thus, the expression $P(X = x)$ symbolizes the probability that the random variable X takes on the particular value x. Often, this is written simply as $P(x)$.
- Likewise, $P(X \leq x)$ = probability that the random variable X is less than or equal to the specific value x;
- $P(a \leq X \leq b)$ = probability that X lies between values a and b.

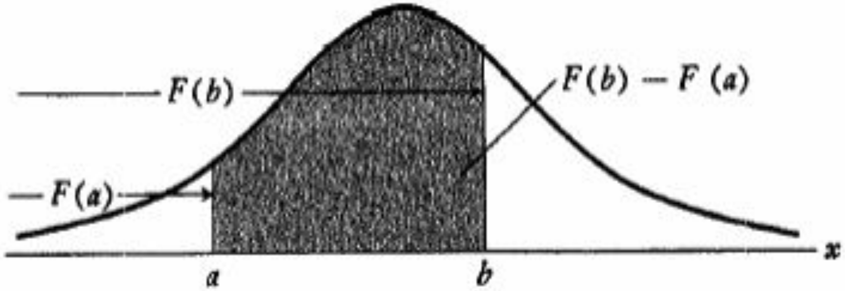
Probability Distribution Function (PDF): specification of the probability associated with each value of a random variable.

For continuous r.v.s:

$$F(a)= p(X \le a)= \int_{-\infty}^{a} f(x)\, dx = \text{Area up to } X = a$$



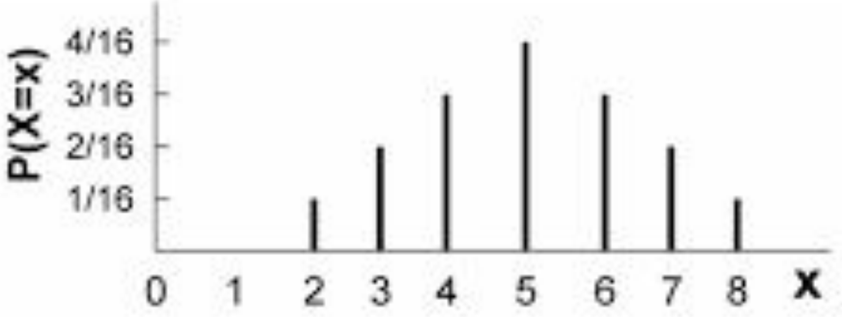
$$p(a \le X \le b) = F(b) - F(a)$$



Probability Mass Function (PMF)

A probability distribution involving only discrete values of X. Aggregates different possible values of X, and the different possible values of P(x).

x	P(x)
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16



Properties:

$0 \leq P(X = x) \leq 1$

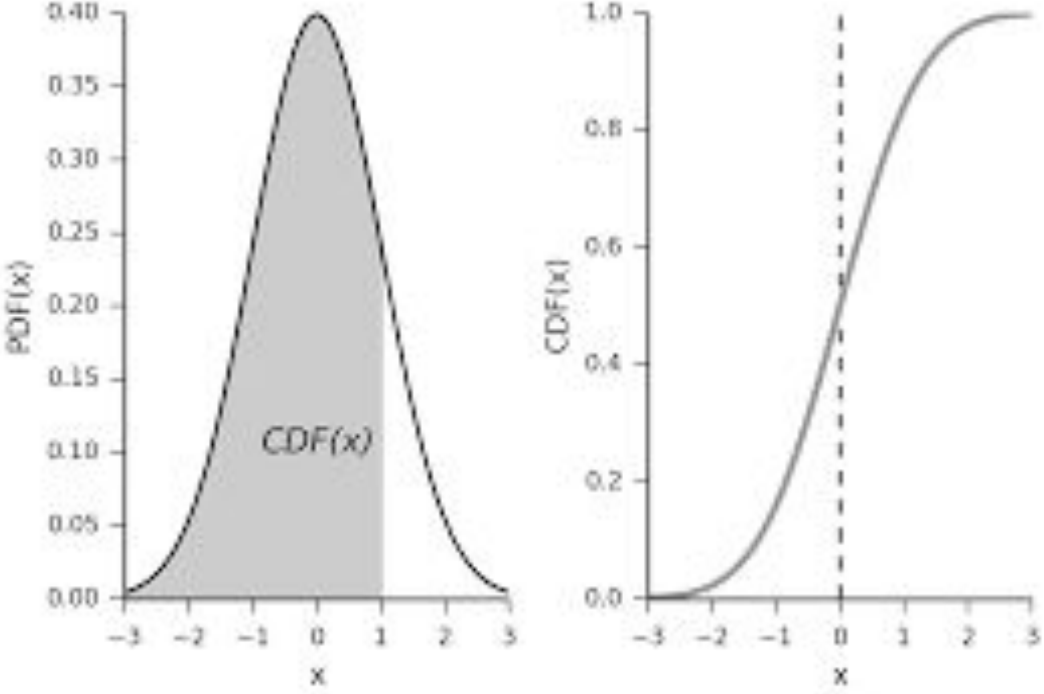
$\Sigma\, P(X = x) = 1.$

Cumulative Density Function (CDF)

The probability that a random variable X takes on a value less than or equal to some particular value a is often written as

$$F(a) = p(X \leq a) = \sum_{X \leq a} p(x)$$

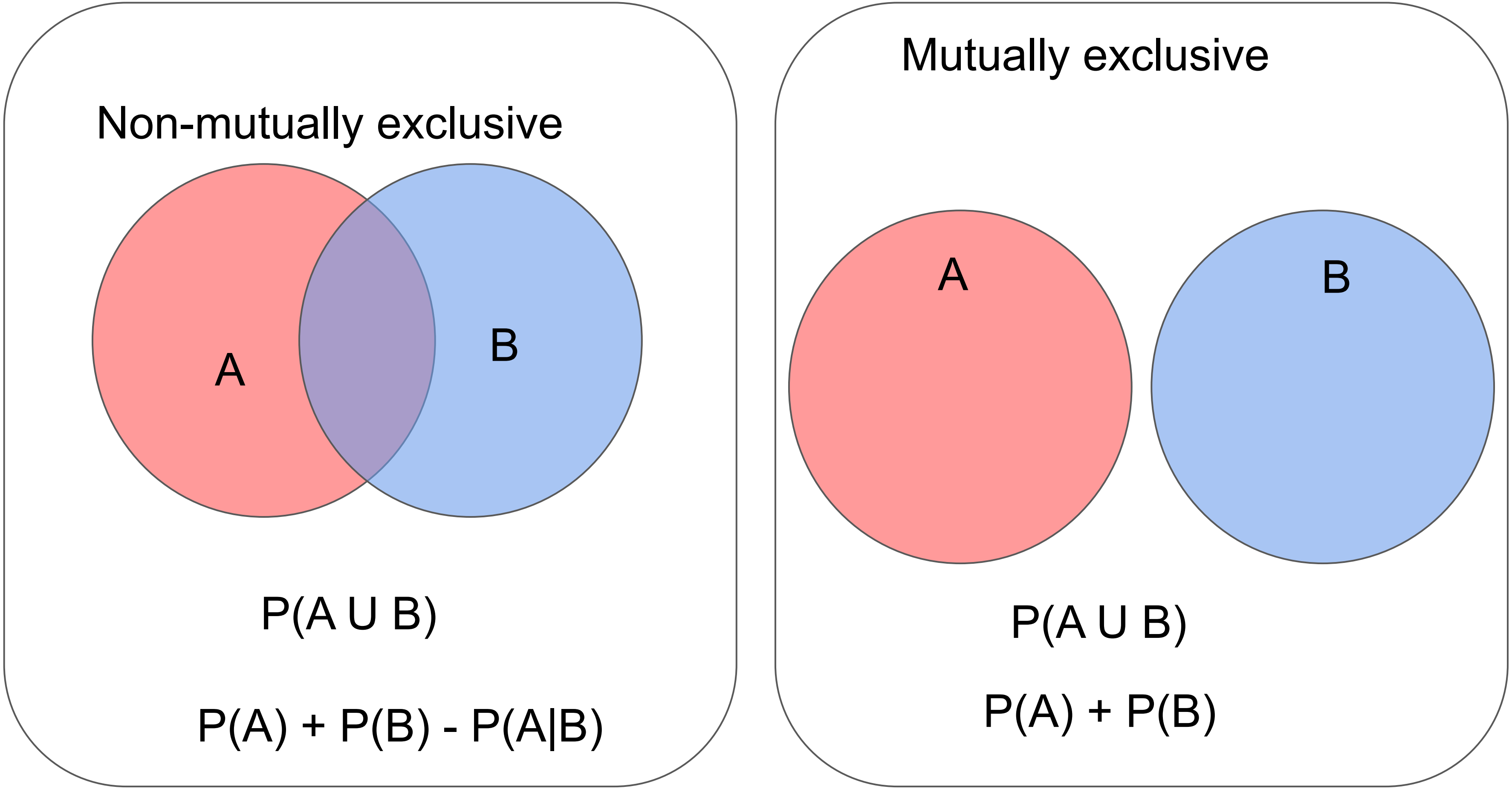
(for discrete variables, integral for continuous)



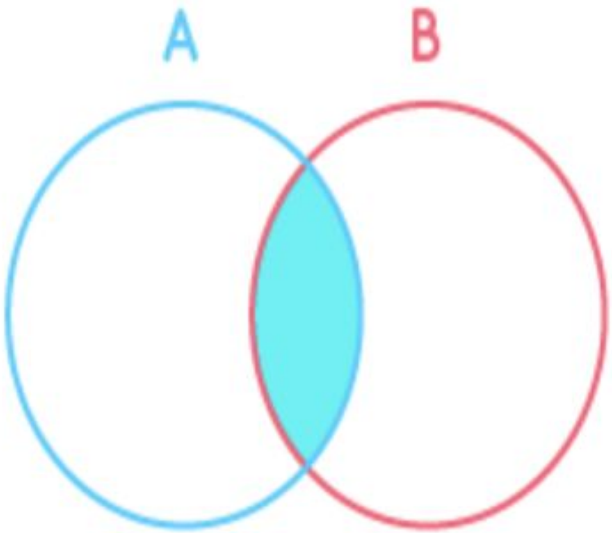
Likelihood of event

$$P(\text{event}) = \frac{\#\text{ of outcomes of event}}{\#\text{ of outcomes in }\Omega}$$

Events


Non-mutually exclusive
A
B
P(A U B)
P(A) + P(B) - P(A|B)
Mutually exclusive
A
B
P(A U B)
P(A) + P(B)

Independent



$P(A \cap B)$

P(A) * P(B)

Summary of probabilities

Event	Probability
A	$P(A) \in [0,1]$
not A	$P(A^{\complement}) = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B) \qquad \text{if A and B are mutually exclusive}$
A and B	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ $P(A \cap B) = P(A)P(B) \qquad \text{if A and B are independent}$
A given B	$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B\|A)P(A)}{P(B)}$

Bayes Rule

- P(A|B) = P(B|A) * P(A) / P(B)
- NOTE: we often do not have access to P(B) and have to calculate by looking at all possible cases:
- P(B) = P(B|A) * P(A) + P(B|not A) * P(not A)

Bayes example 1

At a School, 60% of the boys play football and 36% of the boys play ice hockey.

Given that 40% of those that play football also play ice hockey, what percent of those that play ice hockey also play football?

P(A|B)?

P(A) = 60% = 0.6

P(B) = 36% = 0.36

P(B|A) = 40% = 0.4

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.6 \times 0.4}{0.36} = \frac{0.24}{0.36} = 66\tfrac{2}{3}\%$$

Example 1b

Now, for the problem above, what is the percentage of those that do not play football that play hockey?

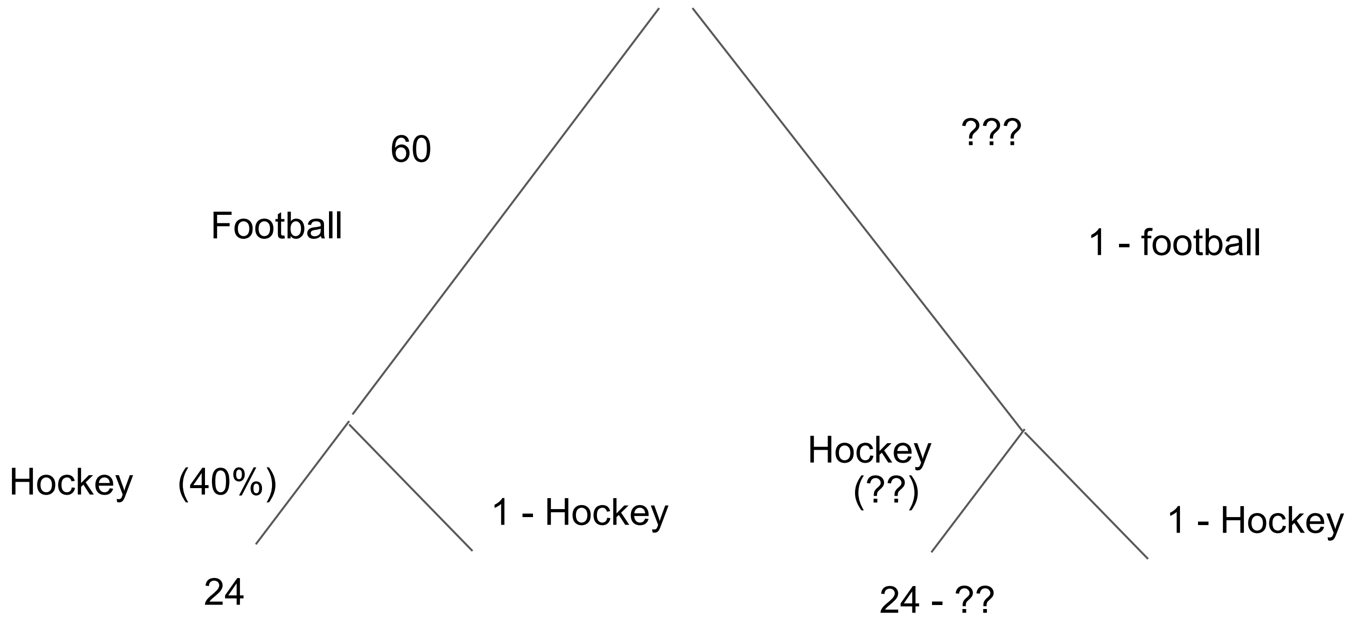
What is the P(H|~F)?

$P(F) = 60\% = 0.6$

$P(H|F) = 40\% = 0.4$

$P(H) = 36\% = P(F)P(H|F) + P(\sim F)P(H|\sim F) = 0.36$

Let’s imagine that there are 100 students…



Bayes Example 4

In a factory, machine X produces 60% of the daily output and machine Y produces 40% of the daily output.

2% of machine X's output is defective, and 1.5% of machine Y's output is defective.

One day, an item was inspected at random and found to be defective. What is the probability that it was produced by machine X?

P(X|defective)

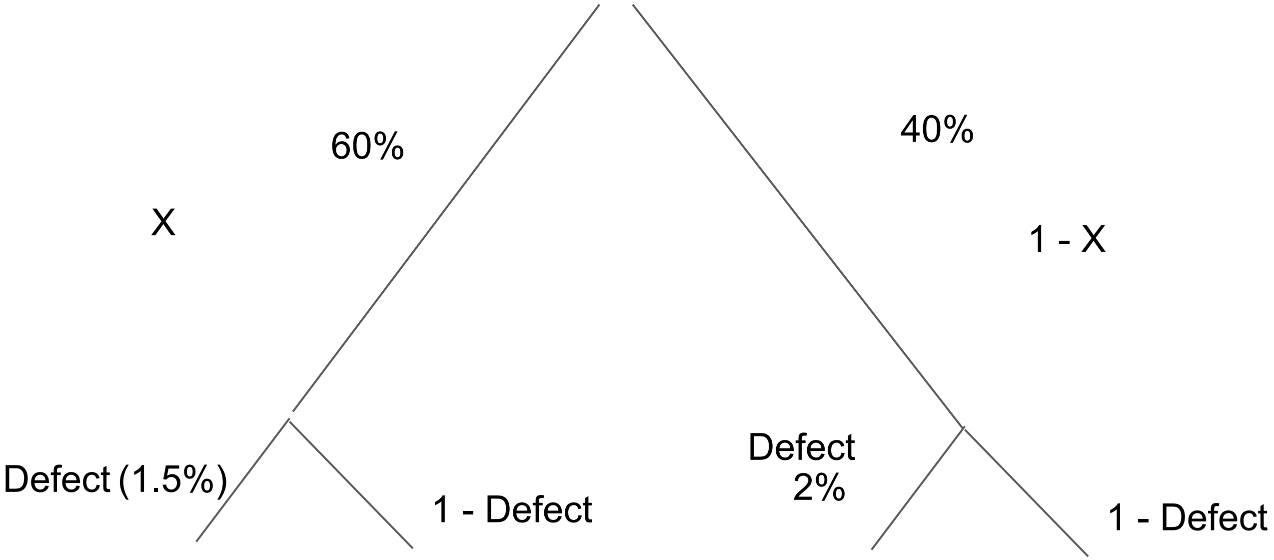
P(X) = 60% = 0.6

P(defective) = 2% × 60% + 1.5% × 40% = 0.012 + 0.006 = 0.018

P(defective|X) = 2% = 0.02

$$\frac{P(A)P(B|A)}{P(B)} = \frac{0.6 \times 0.02}{0.018} = \frac{0.012}{0.018} = \frac{2}{3}$$

Defect



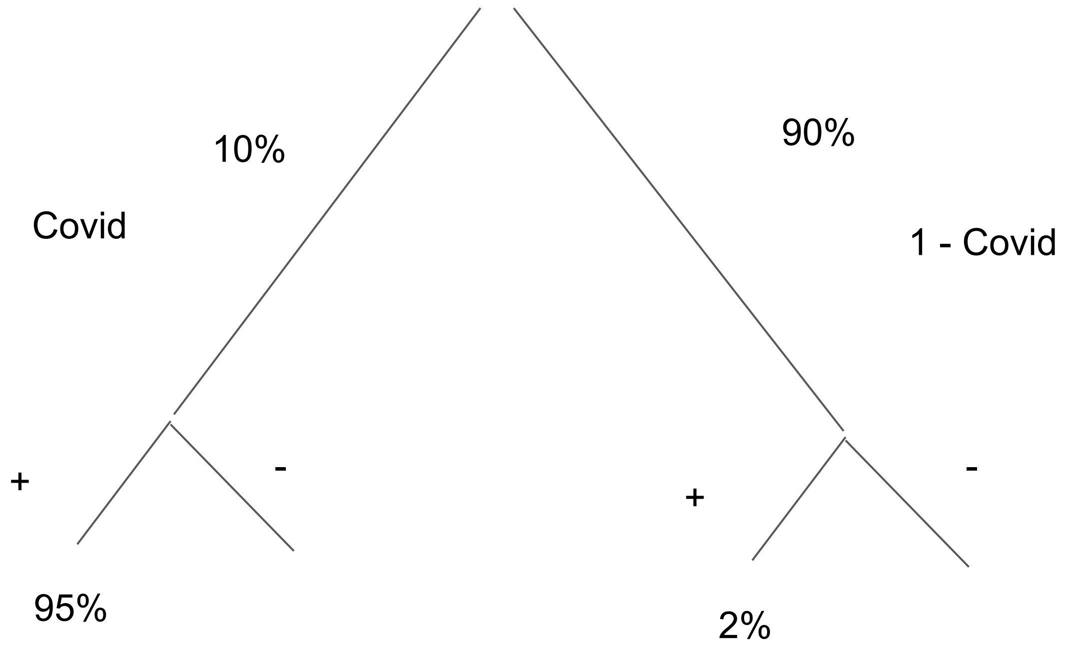
Covid

Let’s 5% of the population has COVID, and test has true positive of 85% (says you have it when you do have it), and false positive of 10% (says you have it when you don’t).

You take a test and it says positive, what is the actual chance that you have it?

Have Covid given a positive test result

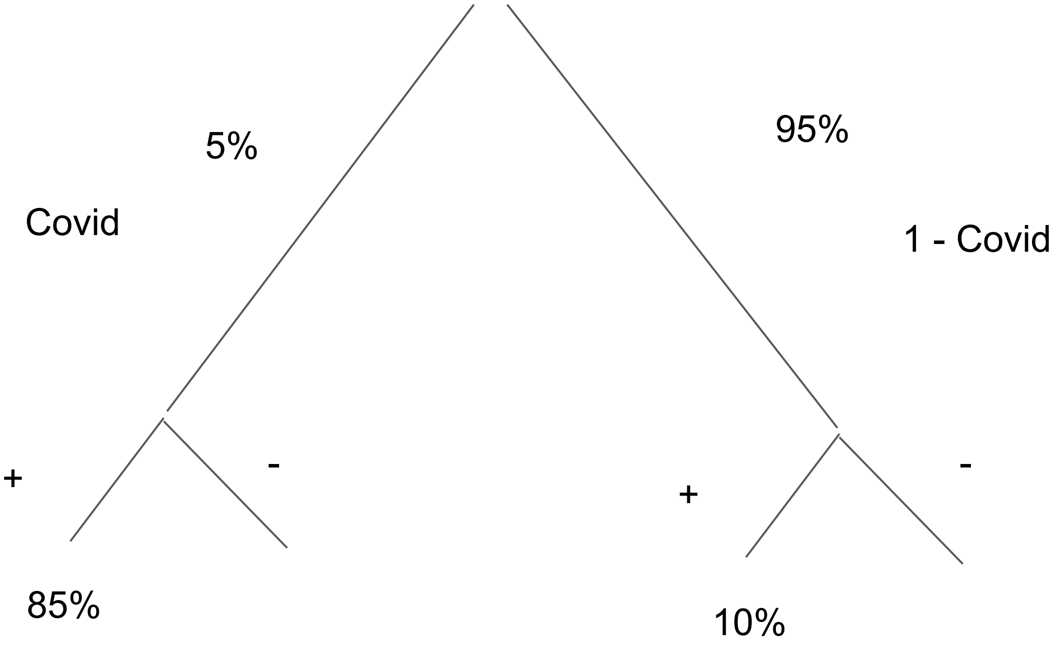
$$\frac{10\% * 95\%}{(10\% * 95\% + 90\% * 2\%)} = 84\%$$



P(B) = P(A)P(B|A) + P(~A)P(B|~A) = 10% * 95% + 90% * 2%

Have Covid given a positive test result

$$\frac{5\% * 85\%}{(5\% * 85\% + 95\% * 10\%)} = 31\%$$



$$P(B) = P(A)P(B|A) + P(\sim A)P(B|\sim A) = 5\% * 85\% + 95\% * 10\%$$

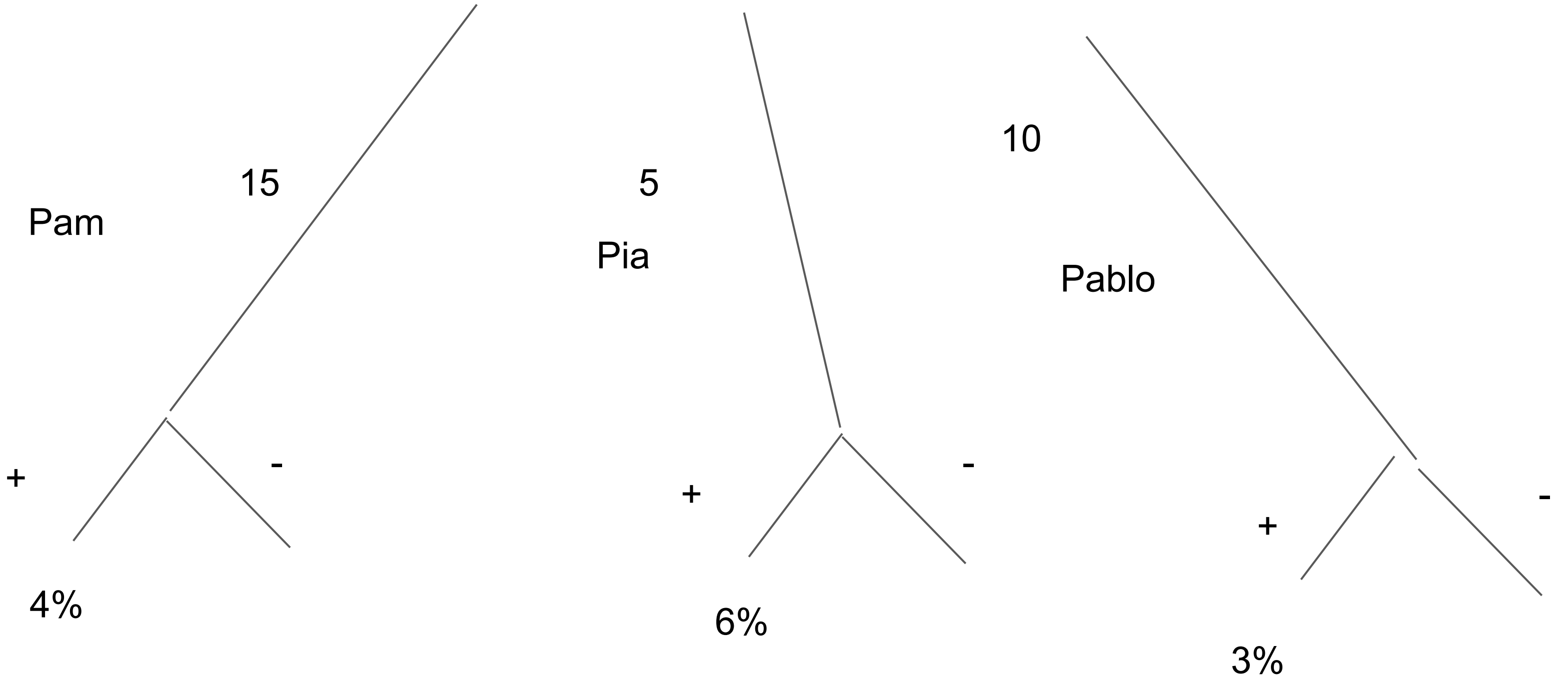
$$P(A1|B) = \frac{P(A1)P(B|A1)}{P(A1)P(B|A1) + P(A2)P(B|A2) + P(A3)P(B|A3) + ...etc}$$

- Pam put in 15 paintings, 4% of her works have won First Prize.
- Pia put in 5 paintings, 6% of her works have won First Prize.
- Pablo put in 10 paintings, 3% of his works have won First Prize.

What is the chance that Pam will win First Prize?

Pam wins?

$$\frac{4\%*15/30}{(4\%*15/30 + 6\%*5/30 + 4\%*10/30)} = 50\%$$



$P(B) = P(A)P(B|A) + P(\sim A)P(B|\sim A) = 4\%*15/30 + 6\%*5/30 + 4\%*10/30$

End of class form

https://forms.gle/kgT2w9wPZo3vJcjA8