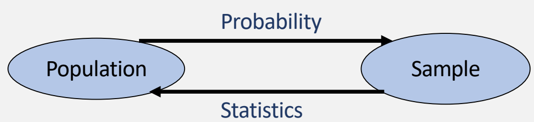
Data Science for Researchers and Scholars

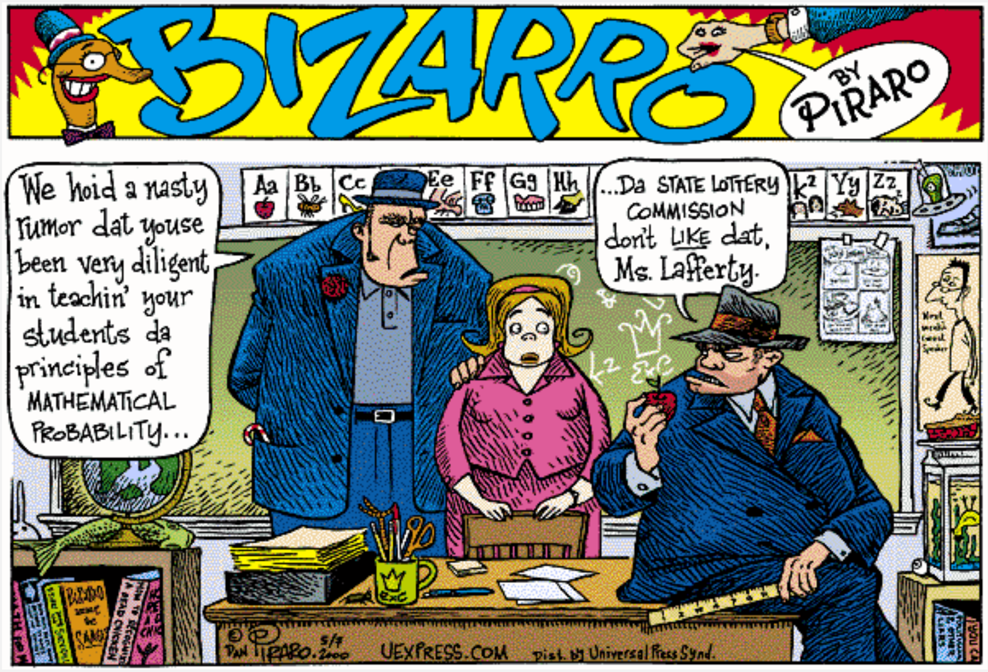
Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Probability

- Why do we care about probability?
  - Nothing in life is certain
  - In everything we do, from drawing inferences from data, to betting on stocks, to assessing a patient's risk of death, we need a means of quantifying uncertainty
  - Probability offers
    - A quantitative measure of the chances associated with various outcomes
    - A bridge between descriptive and inferential statistics
    - A means of making inferences about the population based on what we observe in a sample from the population




BIZARRO
BY PIRARO
We hoid a nasty rumor dat youse been very diligent in teachin' your students da principles of MATHEMATICAL PROBABILITY...
...Da STATE LOTTERY COMMISSION don't LIKE dat, Ms. Lafferty.
UEXPRESS.COM
Dist. by Universal Press Synd.

Probabilistic vs Statistical Reasoning

- Suppose I know exactly the chance that the outcomes of a coin toss.
  - Then I can find the probability that the first toss would be a head.
  - This is probabilistic reasoning as I use knowledge of the population to make predictions about any sample.
- Suppose that I do not know the chances of the two outcomes of a coin toss, but would like to estimate them.
  - I observe a random sample of tosses of the same coin.
  - Suppose I observe n_H heads and n_T tails in a sample of size $n = n_H + n_T$.
  - I estimate of the chance of heads to be $\frac{n_H}{n}$ and of tails to be $\frac{n_T}{n}$.
  - This is statistical reasoning as I am drawing inferences about the population based on what I observe in a sample.

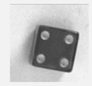
What is Probability?

- In the last lecture, we saw descriptive statistics
- We measured "how often" an outcome of interest is observed in a sample using relative frequency
- For example, the fraction of heads in a sample of coin tosses $\frac{n_H}{n}$
- As the sample size n increases
  - Sample approaches the population
  - Relative frequency of an outcome approaches its probability

Some terminology

- An experiment is the process by which an observation (or measurement) is obtained
- An event is an outcome of an experiment, usually denoted by an uppercase letter
- We associate probabilities with events
  - Outcome of a coin toss
  - The color of a flower in your flower basket
- When an experiment is performed, a particular event either occurs, or it doesn't!
  - The event toss = Head occurs if the coin shows up heads
  - The event color = Red occurs if you happen to pick a red flower

Experiments and Events

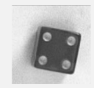


- Experiment: Record an age
  - A: The person is older than 30
  - B: The person is older than 65

  Not mutually exclusive
- Experiment: Toss a die
  - A: The toss comes up odd
  - B: The toss comes up even

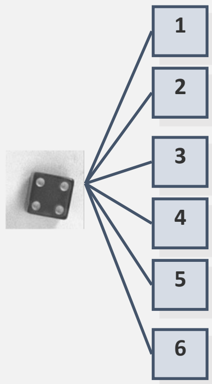
  Mutually exclusive
- Events are
  - mutually exclusive if when one event occurs, the other cannot, and vice versa
  - exhaustive if they cover all possible outcomes
- An event that cannot be decomposed is called a simple event
- Each simple event has a probability associated with it
- Sample space is the set of all simple events (possible outcomes) of an experiment that are mutually exclusive and exhaustive

Example: The 6-sided die toss



Simple events

Sample space



E_1

E_2

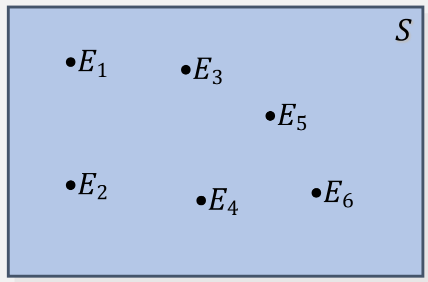
E_3

E_4

E_5

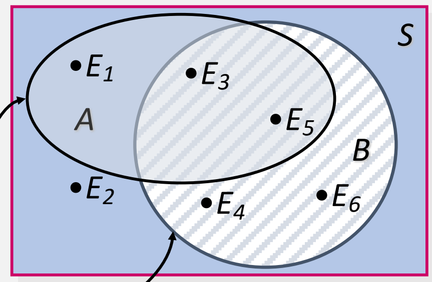
E_6

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$



Some terminology

- An event is a collection of one or more simple events.

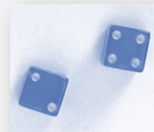


- The die toss
  - A: an odd number
  - B: a number > 2

$$A = \{E_1, E_3, E_5\}$$

$$B = \{E_3, E_4, E_5, E_6\}$$

Probability of an Event



- The probability of an event A measures "how often" A occurs. We denote it by $P(A)$.
- Suppose in an experiment that is performed n times the event A occurs n_A times
- The relative frequency of event A is $\left(\frac{n_A}{n}\right)$
- Then $P(A) = \lim_{n\to\infty}\left(\frac{n_A}{n}\right)$

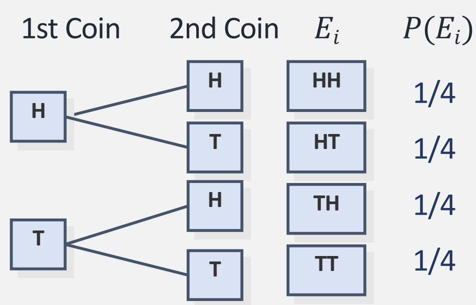
Probability of an Event

- $P(A)$ must be between 0 and 1.
- If event A never occurs*, $P(A) = 0$
- If event A always occurs*, $P(A) = 1$
- The sum of the probabilities for all simple events in S equals 1.
- The probability of an event A is found by adding the probabilities of all the simple events contained in A
- Probabilities are estimated from samples
  - Simplest estimates are relative frequency based
  - More on estimation later

* when the associated experiment is performed

Example

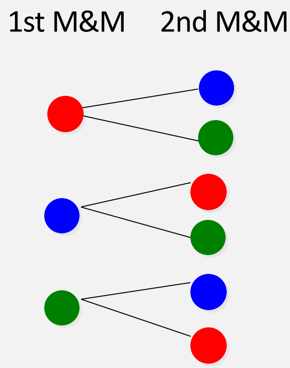
- Suppose we have a fair coin that we toss twice.
- What is the probability of observing at least one head?



$$P(at\ least\ 1\ head) = P(E_1) + P(E_2) + P(E_3) = 1/4 + 1/4 + 1/4 = 3/4$$

Example

- A bowl contains three M&M, one red, one blue and one green.
- A child selects two M&Ms at random.
- What is the probability that at least one is red?

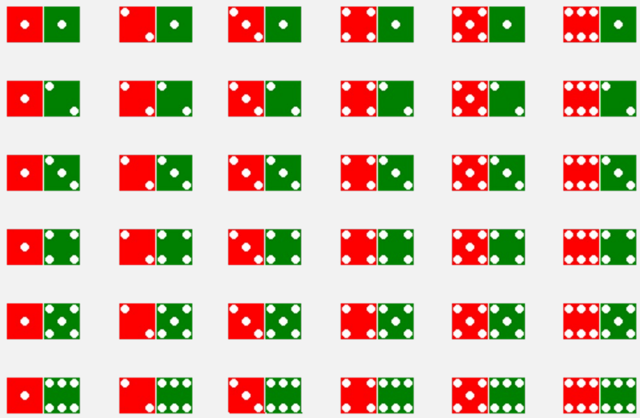


E_i	$P(E_i)$
RB	1/6
RG	1/6
BR	1/6
BG	1/6
GB	1/6
GR	1/6

$$P(at\ least\ 1\ red) = P(RB) + P(BR) + P(RG) + P(GR) = 4/6 = 2/3$$

Example

The sample space of throwing a pair of dice



Example: Throwing a pair of dice

Event	Simple events	Probability
Dice add to 3	(1,2),(2,1)	2/36
Dice add to 6	(1,5),(2,4),(3,3),(4,2),(5,1)	5/36
Red die shows 1	(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)	6/36
Green die shows 1	(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)	6/36

The mn Rule

- If an experiment is performed in two stages, with m ways to accomplish the first stage and n ways to accomplish the second stage, then there are mn ways to accomplish the experiment.
- This rule is easily extended to k stages, with the number of ways equal to $n_1 n_2 n_3 \ \dots \ n_k$

Example:

- Toss two coins.
- The total number of simple events is: $2 \times 2 = 4$

Examples

Toss three coins. The total number of simple events is:

$$2 \times 2 \times 2 = 8$$

Toss two dice. The total number of simple events is:

$$6 \times 6 = 36$$

Toss three dice. The total number of simple events is:

$$6 \times 6 \times 6 = 216$$

Two M&Ms are drawn from a dish containing two red and two blue candies. The total number of simple events is:

$$4 \times 3 = 12$$

Counting events: Permutations

- The number of ways you can arrange n distinct objects, taking them r at a time is

$$P_r^n = \frac{n!}{(n-r)!}$$

where $n! = n(n-1)(n-2)\ldots(2)(1)$ and $0! \equiv 1$.

How many 3-digit lock combinations can we make from the numbers 1, 2, 3, and 4?

The order of the choice is important!

$$P_3^4 = \frac{4!}{1!} = 4(3)(2) = 24$$

Combinations

- The number of distinct combinations of n distinct objects that can be formed, taking them r at a time is

$$C_r^n = \frac{n!}{r!(n-r)!}$$

- Three members of a 5-person committee must be chosen to form a subcommittee.
- How many different subcommittee compositions are there?

The order of the choice is *not* important!

$$C_3^5 = \frac{5!}{3!(5-3)!} = \frac{5(4)(3)(2)1}{3(2)(1)(2)1} = \frac{5(4)}{(2)1} = 10$$

Example

- A box contains six M&Ms, four red and two green.
- A child selects two M&Ms at random.
- What is the probability that exactly one is red?

The order of the choice is not important!

$$C_2^6 = \frac{6!}{2!4!} = \frac{6(5)}{2(1)} = 15$$
ways to choose 2 M & Ms.

$$C_1^2 = \frac{2!}{1!1!} = 2$$
ways to choose 1 green M & M.

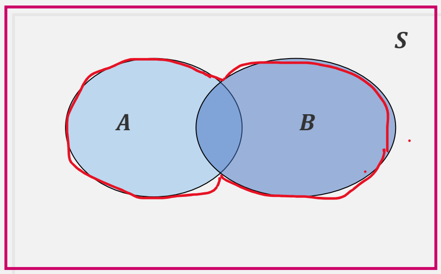
$$C_1^4 = \frac{4!}{1!3!} = 4$$
ways to choose 1 red M & M.

4×2 =8 ways to choose 1 red and 1 green M&M.

P(exactly one red) = 8/15

Event Relations

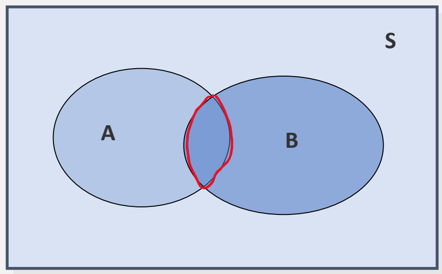
- We can combine events to make other events using logical operations: and, or and not.
- The union of two events, A and B, is the event that either A or B or both occur when the experiment is performed.
- We write $A \cup B$



Event Relations

- The **intersection** of two events, A and $\boldsymbol{B}$, is the event that both A **and** B occur when the experiment is performed. We write $A \cap B$.

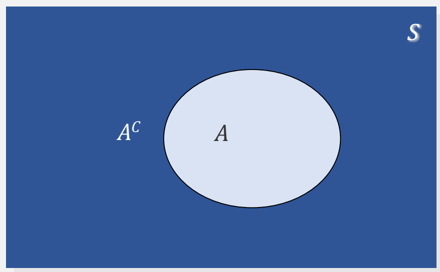
$A \cap B$



- If two events A and B are mutually exclusive, then $P(A \cap B) = 0$

Event Relations

- $\boldsymbol{A^C}$, the **complement** of an event $\boldsymbol{A}$ consists of all outcomes of the experiment that do not result in event $\boldsymbol{A}$**.**



Example

- Select a student from a classroom with only male or female students and record the student's hair color and gender.
  - A: student has brown hair
  - B: student is female
  - C: student is male

Mutually exclusive; $B = C^C$

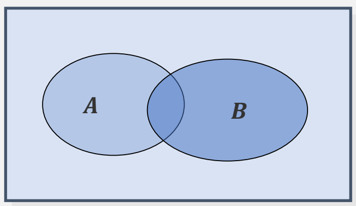
What is the relationship between events B and C?

- A^C: Student does not have brown hair
- $B \cap C$: Student is both male and female = $\varnothing$
- $B \cup C$: Student is either male and female = all students = S

Probabilities of unions

- For any two events, A and B, the probability of their union, $P(A \cup B)$, is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Example: Additive Rule

Example: Suppose that there were 120 students in the classroom, and that they could be classified as follows

A: brown hair
$P(A) = 50/120$
B: female
$P(B) = 60/120$

	Brown	Not Brown
Male	20	40
Female	30	30

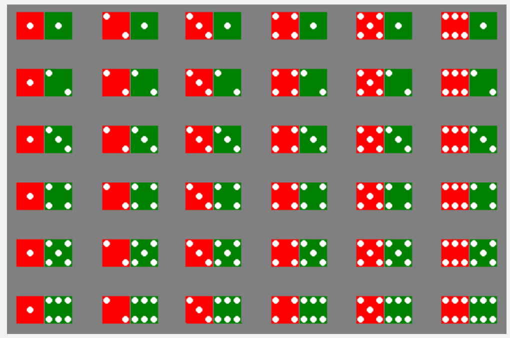
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 50/120 + 60/120 - 30/120$$
$$= 80/120 = 2/3$$

Check: $P(A \cup B)$
$= (20 + 30 + 30)/120$

Example: Two Dice

A: red die shows 1

B: green die shows 1



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 6/36 + 6/36 - 1/36$$
$$= 11/36$$

A Special Case

- When two events A and B are mutually exclusive,
- $P(A \cap B) = 0$ and $P(A \cup B) = P(A) + P(B)$.

A: male with brown hair
$P(A) = 20/120$
B: female with brown hair
$P(B) = 30/120$

	Brown	Not Brown
Male	20	40
Female	30	30

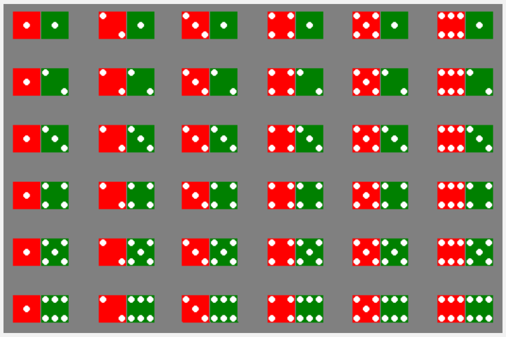
A and B are mutually exclusive, so that

$$P(A \cup B) = P(A) + P(B) = 20/120 + 30/120 = 50/120$$

Example: Two Dice

A: dice add to 3

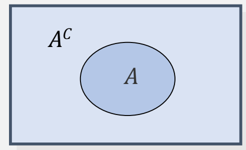
B: dice add to 6



A and B are mutually exclusive, so that

$$P(A \cup B) = P(A) + P(B) = 2/36 + 5/36 = 7/36$$

Probabilities of Complements



- We know that for any event A:
  $P(A \cap A^C) = 0$
- Since either A or A^C must occur,
  $P(A \cup A^C) = 1$
- so that $P(A \cup A^C) = P(A) + P(A^C) = 1$

$$P(A^C) = 1 - P(A)$$

Example

- Select a student at random from the classroom.

A: male
P(A) = 60/120
B: female
P(B) = ?

	Brown	Not Brown
Male	20	40
Female	30	30

A and B are complementary, so that

$$P(B) = 1 - P(A) = 1 - 60/120 = 60/120$$

Independence

- Two events, A and B, are said to be independent if the occurrence or nonoccurrence of one of the events does not change the probability of the occurrence of the other event.
- Example: The color of my shirt being blue and whether the Intel stock price goes up

Conditional Probabilities

The probability that A occurs, given that event B has occurred is called the conditional probability of A given B (written $P(A|B)$)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0$$

Example: Conditional probability

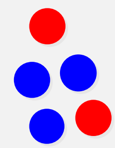
- Toss a fair coin twice.
- The tosses are independent
- $P(A|B) = ½ = P(A|B^c)$
  - A: head on second toss
  - B: head on first toss

HH	1/4
HT	1/4
TH	1/4
TT	1/4

P(A) does not change, whether B happens or not… ⟶ A and B are independent!

Example 2

- A bowl contains five M&Ms, two red and three blue.
- Randomly select two candies
  - A: second candy is red.
  - B: first candy is blue.



$$P(A|B) = P(2^{nd}\, red|1^{st}\, blue) = 2/4 = 1/2$$
$$P(A|\, B^c) = P(2^{nd}\, red|1^{st}\, red) = 1/4$$

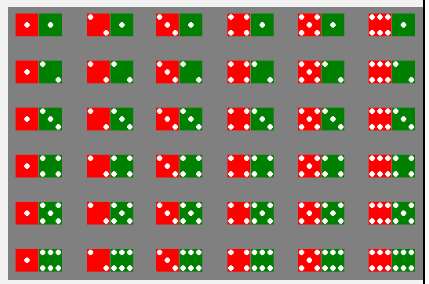
$P(A)$ does change, depending on whether B happens or not… → A and B are dependent!

Exercise: Two Dice

Toss a pair of fair dice.

- A: red die shows 1
- B: green die shows 1

Are A and B independent?



Defining Independence

- We can redefine independence in terms of conditional probabilities
- Two events A and B are independent if and only if
  - $P(A|B) = P(A)$ or $P(B|A) = P(B)$

The Multiplicative Rule for Intersections

- For any two events, A and B, the probability that both A and B occur is $P(A \cap B) = P(A)P(B|A)$
- If A and B are independent then $P(A \cap B) = P(A)P(B)$

Example

- In a certain population, 10% of the people can be classified as being high risk for a heart attack.
- Three people are randomly selected from this population.
- What is the probability that exactly one of the three are high risk?

H: high risk N: not high risk

$$
\begin{aligned}
P(\mathit{exactly\ one\ high\ risk}) &= P(HNN) + P(NHN) + P(NNH) \\
&= P(H)P(N)P(N) + P(N)P(H)P(N) + P(N)P(N)P(H) \\
&= (.1)(.9)(.9) + (.9)(.1)(.9) + (.9)(.9)(.1) \\
&= 3(.1)(.9)^2 = .243
\end{aligned}
$$

Example

- Suppose we know that 49% of the population are female.
- Also, of the female patients, 8% are high risk.
- A single person is selected at random.
- What is the probability that it is a high risk female?

H: high risk F: female

$P(F) = .49$ and $P(H|F) = .08$.

From the Multiplicative Rule,

$$\begin{aligned} P(high\,risk\,female) &= P(H \cap F) \\ &= P(F)P(H|F) \\ &= .49(.08) \\ &= .0392 \end{aligned}$$

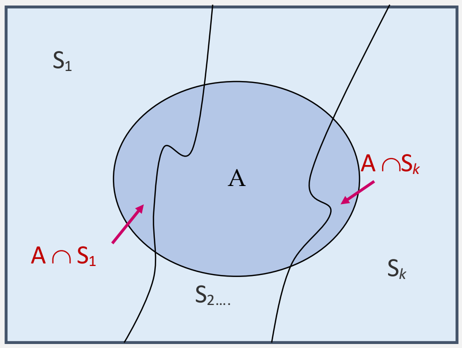
The Law of Total Probability

- Let S_1, S_2, S_3, …, S_k be mutually exclusive and exhaustive events (that is, one and only one can occur).
- Then the probability of any event A can be written as

$$P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$$

$$= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots + P(S_k)P(A|S_k)$$

The Law of Total Probability



$$P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$$
$$= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots$$
$$+ P(S_k)P(A|S_k)$$

Bayes' Rule

Let $S_1, S_2, S_3, \ldots, S_k$ be mutually exclusive and exhaustive events with prior probabilities $P(S_1), P(S_2), \ldots, P(S_k)$. If an event A occurs, the posterior probability of S_i, given that A occurred is

$$P(S_i \mid A) = \frac{P(S_i)P(A \mid S_i)}{\sum P(S_i)P(A \mid S_i)} \text{ for } i = 1, 2, \ldots k$$

Example

- 49% of the population are female.
- Of the female patients, 8% are high risk for heart attack, while 12% of the male patients are high risk.
- A single person is selected at random and found to be high risk.
- What is the probability that it is a male?

H: high risk F: female M: male

$$P(F) = 0.49$$
$$P(M) = 0.51$$
$$P(H|F) = 0.08$$
$$P(H|M) = 0.12$$

$$P(M \mid H) = \frac{P(M)P(H \mid M)}{P(M)P(H \mid M) + P(F)P(H \mid F)}$$
$$= \frac{.51(.12)}{.51(.12) + .49(.08)} = .61$$

Exercise

- Suppose a rare disease infects one out of every 1000 people in a population.
- Suppose that there is a good, but not perfect, test for this disease
  - if a person has the disease, the test comes back positive 99% of the time.
  - On the other hand, the test also produces some false positives: 2% of uninfected people are also test positive.
- And someone just tested positive.
- What are his chances of having this disease?

Example

A survey of job satisfaction[2] of teachers gave the following results

		Job Satisfaction		
		Satisfied	Unsatisfied	Total
LEVEL	College	74	43	117
	High School	224	171	395
	Elementary	126	140	266
	Total	424	354	778

2 "Psychology of the Scientist: Work Related Attitudes of U.S. Scientists" (*Psychological Reports* (1991): 443 – 450).

Example

- If each cell is divided by the total number surveyed, 778, the resulting table is a table of estimated probabilities.

LEVEL	Job Satisfaction: Satisfied	Job Satisfaction: Unsatisfied	Total
College	0.095	0.055	0.150
High School	0.288	0.220	0.508
Elementary	0.162	0.180	0.342
Total	0.545	0.455	1.000

Let $S = \text{Satisfied}$, $C = \text{College}$ .....

- $P(C) = 0.15$ (proportion of teachers who are college teachers)
- $P(S) = 0.545$ (proportion of teachers who are satisfied with their jobs)
- $P(S \cap C) = 0.095$ (proportion of teachers who are college teachers and are satisfied with their jobs)

Example

		Job Satisfaction		
		Satisfied	Unsatisfied	Total
LEVEL	College	0.095	0.055	0.150
	High School	0.288	0.220	0.508
	Elementary	0.162	0.180	0.342
	Total	0.545	0.455	1.000

$$P(C|S) = \frac{P(C \cap S)}{P(S)} = \frac{0.095}{0.545} = 0.175$$

This is the proportion of satisfied teachers that are college teachers.

$$P(S|C) = \frac{P(C \cap S)}{P(C)} = \frac{0.095}{0.150} = 0.632$$

This is the proportion of college teachers that are satisfied.

Example

LEVEL		Job Satisfaction: Satisfied	Unsatisfied	Total
L E V E L	College	0.095	0.055	0.150
	High School	0.288	0.220	0.508
	Elementary	0.162	0.180	0.342
	Total	0.545	0.455	1.000

Are C and S independent events?

$$P(C\ ) \;=\; 0.150$$
$$P(C|S) \;=\; 0.175$$

So clearly, C and S are NOT independent

Exercise

- Tom and Jane are going to take a driver's test at the nearest DMV office.
- Tom estimates that his chances to pass the test are 70% and Jane estimates her chances of passing as 80%.
- Tom and Jane take their tests independently.
- What is the probability that at least one of the two friends pass the test?
- Suppose we learn that only one of the two friends passed the test. What is the probability that it was Jane?

Random Variables

- A variable x is a random variable if the value that it assumes, corresponding to the outcome of an experiment is a chance or random event.
- Random variables can be discrete or continuous

Examples

- x = SAT score for a randomly selected student
- x = number of people who click on your website on a randomly chosen of the year 2023
- x = outcome of a die toss

Probability Distributions of Discrete Random Variables

- The probability distribution for a discrete random variable x is a graph, table or formula that gives the probability $p(x)$ associated with each value of x
- Note that
  - $\forall x \ 0 \leq p(x) \leq 1$
  - $\sum_i \ p(x = v_i) = 1$ (*i* indexes over the possible values v_i of x

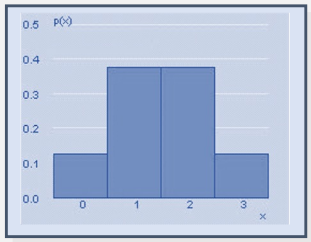
Example

Toss a fair coin three times and define x = number of heads.

		x
HHH	1/8	3
HHT	1/8	2
HTH	1/8	2
THH	1/8	2
HTT	1/8	1
THT	1/8	1
TTH	1/8	1
TTT	1/8	0

$$P(x=0) = 1/8$$
$$P(x=1) = 3/8$$
$$P(x=2) = 3/8$$
$$P(x=3) = 1/8$$

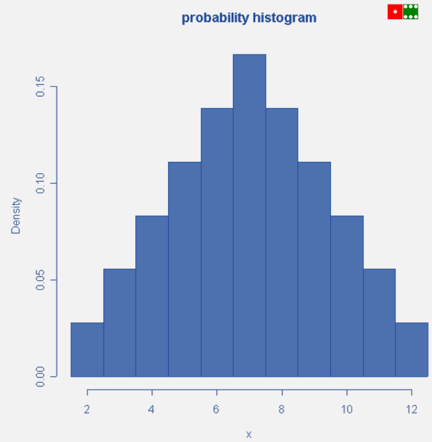
x	$p(x)$
0	1/8
1	3/8
2	3/8
3	1/8



Probability distribution of x

Example

Toss two dice and define
x = sum of two dice.



x	$p(x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Probability Distributions

- Probability distributions can be used to describe the population, just as we described samples using statistics
- Shape: Symmetric, skewed, mound-shaped…
  - Outliers: unusual or unlikely measurements
  - Center and spread: mean and standard deviation. A population mean is called μ and a population standard deviation is called σ.
- Let x be a discrete random variable with probability distribution $p(x)$. Then the mean, variance and standard deviation of x are given as

$$\text{Mean}: \mu = \sum x p(x)$$

$$\text{Variance}: \sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\text{Standard deviation}: \sigma = \sqrt{\sigma^2}$$

Example

Toss a fair coin 3 times and record x, the number of heads.

x	$p(x)$	$xp(x)$	$(x-\mu)^2 p(x)$
0	1/8	0	$(-1.5)^2(1/8)$
1	3/8	3/8	$(-0.5)^2(3/8)$
2	3/8	6/8	$(0.5)^2(3/8)$
3	1/8	3/8	$(1.5)^2(1/8)$

$$\mu = \sum xp(x) = \frac{12}{8} = 1.5$$

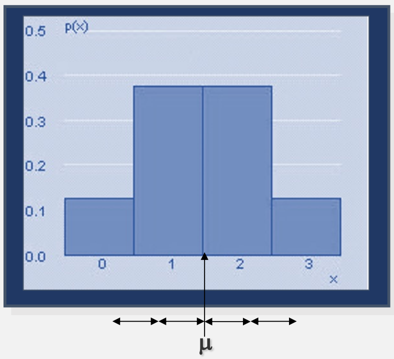
$$\sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\sigma^2 = .28125 + .09375 + .09375 + .28125 = .75$$

$$\sigma = \sqrt{.75} = .688$$

Example

The probability distribution for x, the number of heads in tossing 3 fair coins.



- Shape? – Symmetric
- Outliers? - None
- Center? $\mu = 1.5$
- Spread? $\sigma = .688$