19CSE304 - FOUNDATIONS OF DATA SCIENCE

FDS_02: Introduction to Data Science

Dr. Venugopal K

Assoc Professor, CSE, ASE, Amritapuri

Conditional Probability

- marginal, joint, conditional probability
- Bayes' theorem
- general product rule

Conditional Probability

CONDITIONAL PROBABILITY

The conditional probability of outcome A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Conditional Probability

- If two events A and B are not independent, then information available about the outcome of event A can influence the predictability of event B
- Conditional probability
  - $P(B \mid A) = P(A \cap B)/P(A)$ if $P(A) > 0$
  - $P(A \mid B)P(B) = P(B \mid A)P(A)$ - Bayes formula
  - $P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$
- Example: two (fair) coin toss experiment
  - Event A : First toss is head = {HT, HH}
  - Event B : Two successive heads ={HH}
  - Pr(B) = 0.25 (no information)
  - Given event A has occurred $Pr(B|A) = 0.5 = 0.25/0.5 = P(A \cap B)/P(A)$

Probability distributions

A probability distribution is a table of all disjoint outcomes and their associated probabilities.

one toss	head	tail
probability	0.5	0.5

two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

Rules for probability distributions:

1. The events listed must be disjoint
2. Each probability must be between 0 and 1
3. The probabilities must total 1

General multiplication rule

Product rule for independent events:

If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

General product rule:

$P(A \text{ and } B) = P(A \mid B) \times P(B)$

General multiplication rule

- Earlier we saw that if two events are independent, their joint probability is simply the product of their probabilities. *If the events are not believed to be independent, the joint probability is calculated slightly differently.*
- If A and B represent two outcomes or events, then

$$\mathbf{P(A \text{ and } B) = P(A \mid B) \times P(B)}$$

Note that this formula is simply the conditional probability formula, rearranged.

- It is useful to think of A as the outcome of interest and B as the condition.

Independence and conditional probabilities

Generically, if $P(A|B) = P(A)$ then the events A and B are said to be independent.

- Conceptually: Giving B doesn't tell us anything about A.
- Mathematically: If events A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$. Then,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Independence and conditional probabilities

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is 60 / 100 = 0.6.
- The probability that a randomly selected student is a social science major given that they are female is 30 / 50 = 0.6.
- Since *P(SS | M)* also equals 0.6, major of students in this class does not depend on their gender: *P(SS | F) = P(SS)*.

Example: Smallpox in Boston, 1721

The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston. Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: **inoculated** and **result**. The variable inoculated takes two levels: **yes** or **no**, indicating whether the person was inoculated or not. The variable result has outcomes lived or died. These data are summarized in Tables A and B.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Figure A: Contingency table for the smallpox data set.

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Figure B: Table proportions for the smallpox data, computed by dividing each count by the table total, 6224

Sample Problem

Q1. Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability

Q2. Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of the answer to Q1.

$$P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{no}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{no})}{P(\texttt{inoculated} = \texttt{no})} = \frac{0.1356}{0.9608} = 0.1411.$$

$$P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{yes}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{yes})}{P(\texttt{inoculated} = \texttt{yes})} = \frac{0.0010}{0.0392} = 0.0255$$

Tree Diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

Ex1: smallpox data

In the smallpox data, we see the population as split by **inoculation**: **yes** and **no**. Following this split, survival rates were observed for each group. This structure is reflected in the tree diagram shown in Figure C.

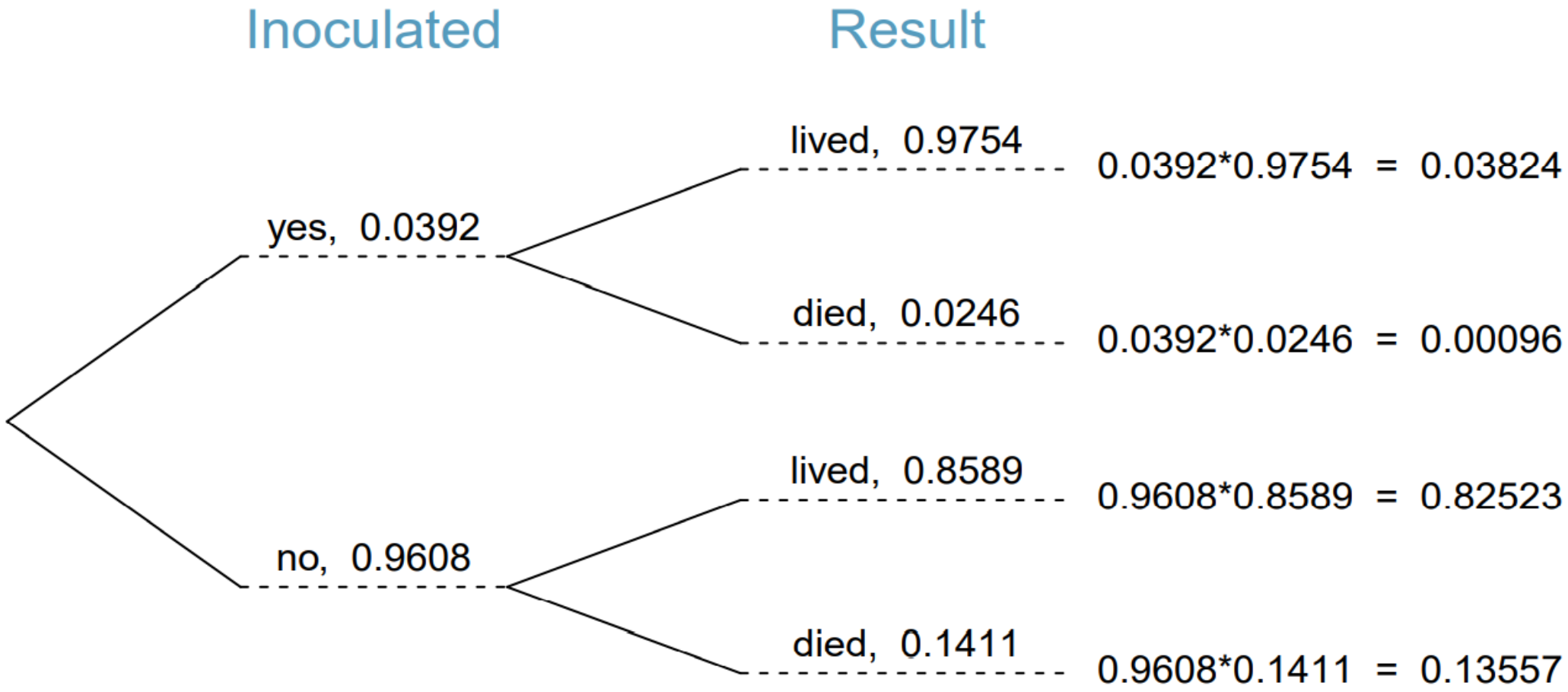


Figure C: A tree diagram of the smallpox data set

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure C. This tree diagram splits the smallpox data by inoculation into the yes and no groups with respective marginal probabilities **0.0392** and **0.9608**.

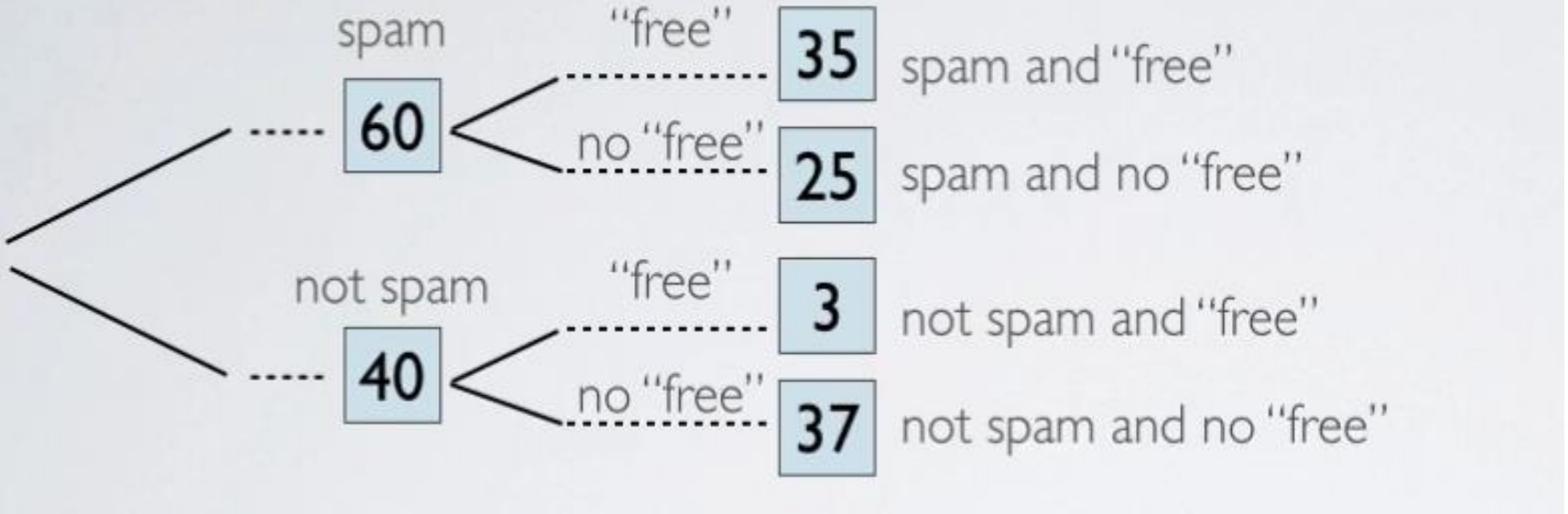
The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure C is the probability that **result = lived** conditioned on the information that **inoculated = yes**.

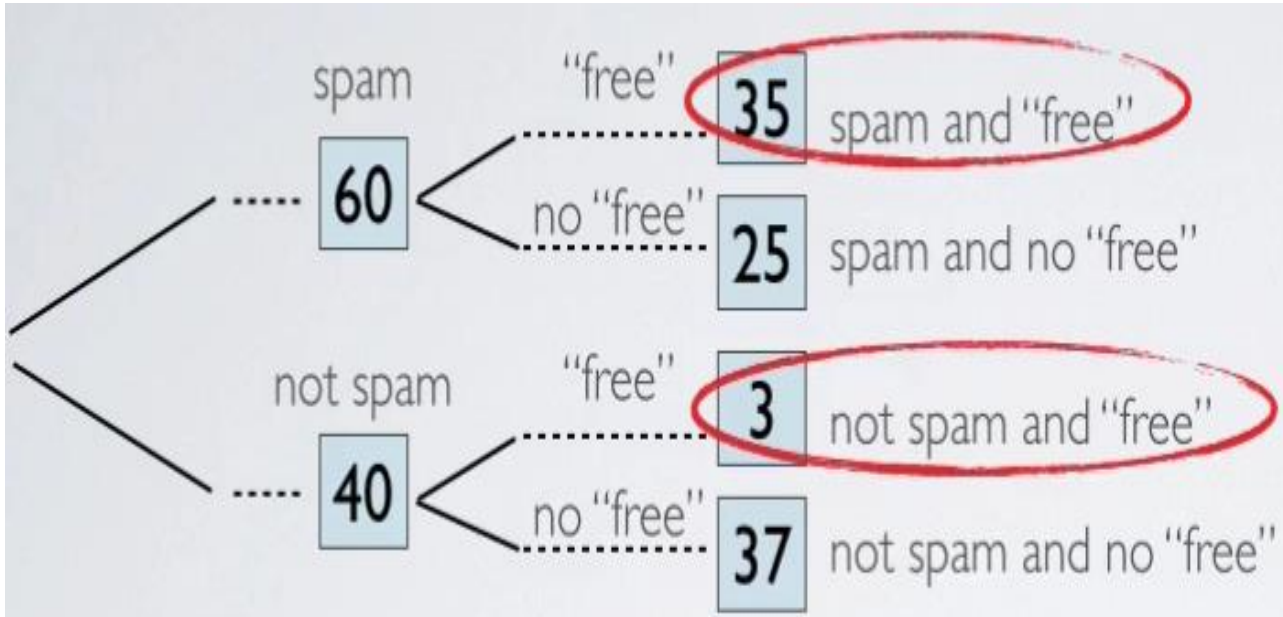
Joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned} &P(\texttt{inoculated} = \texttt{yes} \text{ and } \texttt{result} = \texttt{lived}) \\ &\quad = P(\texttt{inoculated} = \texttt{yes}) \times P(\texttt{result} = \texttt{lived} | \texttt{inoculated} = \texttt{yes}) \\ &\quad = 0.0392 \times 0.9754 = 0.0382 \end{aligned}$$

Ex2: Probability of Spam mail

You have 100 emails in your inbox: 60 are spam, 40 are not. Of the 60 spam emails, 35 contain the word "free". Of the rest, 3 contain the word "free". If an email contains the word "free", what is the probability that it is spam?





$$P(\text{spam} \mid \text{"free"}) = \frac{35}{35 + 3} = 0.92$$

Bayesian Inference example 1

Breast cancer screening

- American Cancer Society estimates that about **1.7%** of women have breast cancer. *http://www.cancer.org/cancer/cancerbasics/cancer-prevalence*
- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about **78%** of women who truly have breast cancer. *http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html*
- An article published in 2003 suggests that up to **10%** of all mammograms result in **false positives** for patients who do not have cancer. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940*

Note: These percentages are approximate, and very difficult to estimate.

American Cancer Society estimates that about 1.7% of women have breast cancer.

http:// www.cancer.org/ cancer/ cancerbasics/ cancer-prevalence

Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.

http:// ww5.komen.org/ BreastCancer/ AccuracyofMammograms.html

An article published in 2003 suggests that up to 10% of all mammograms are false positive.

http:// www.ncbi.nlm.nih.gov/ pmc/ articles/ PMC1360940

$P(bc) = 0.017$

$P(+ \mid bc) = 0.78$

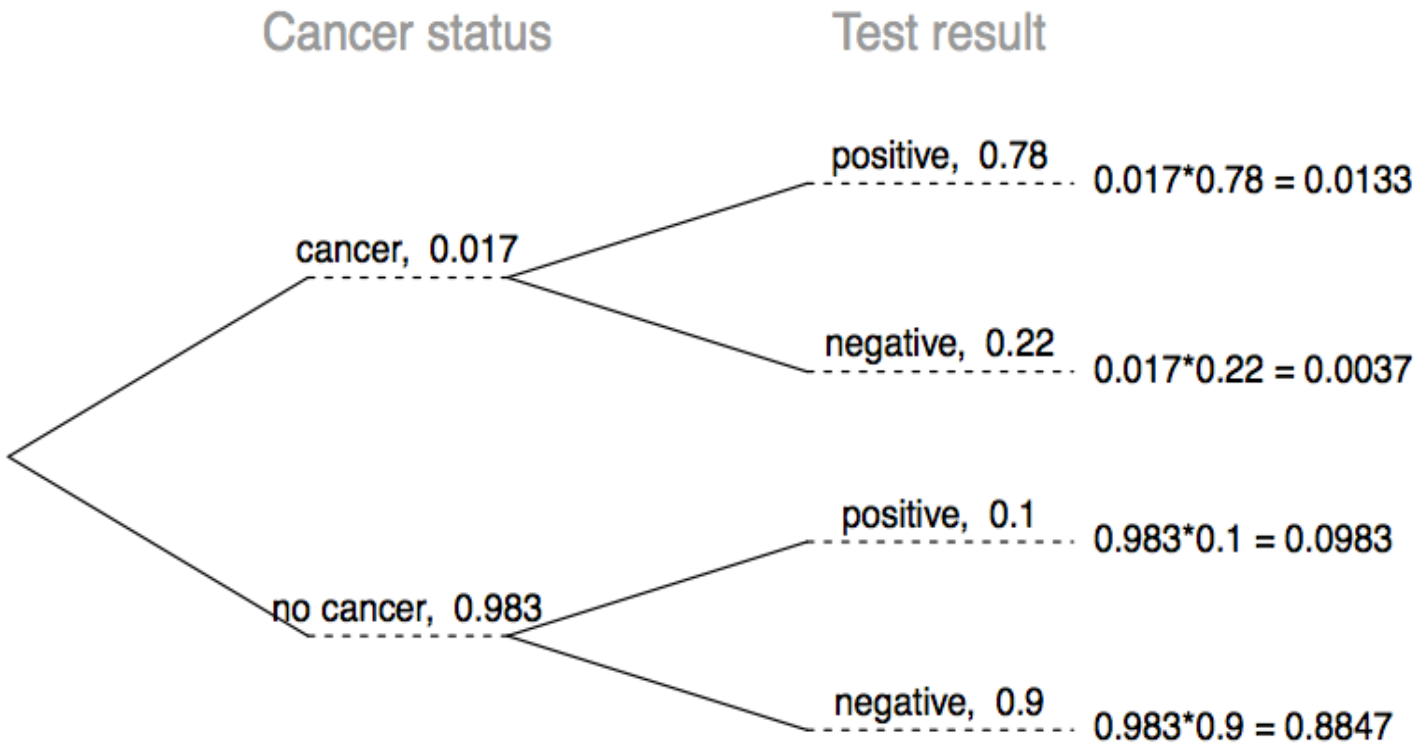
$P(+ \mid \text{no } bc) = 0.10$

Prior to any testing and any information exchange between the patient and the doctor, what probability should a doctor assign to a female patient having breast cancer?

$P(bc) = 0.017$ → prior

Inverting probabilities

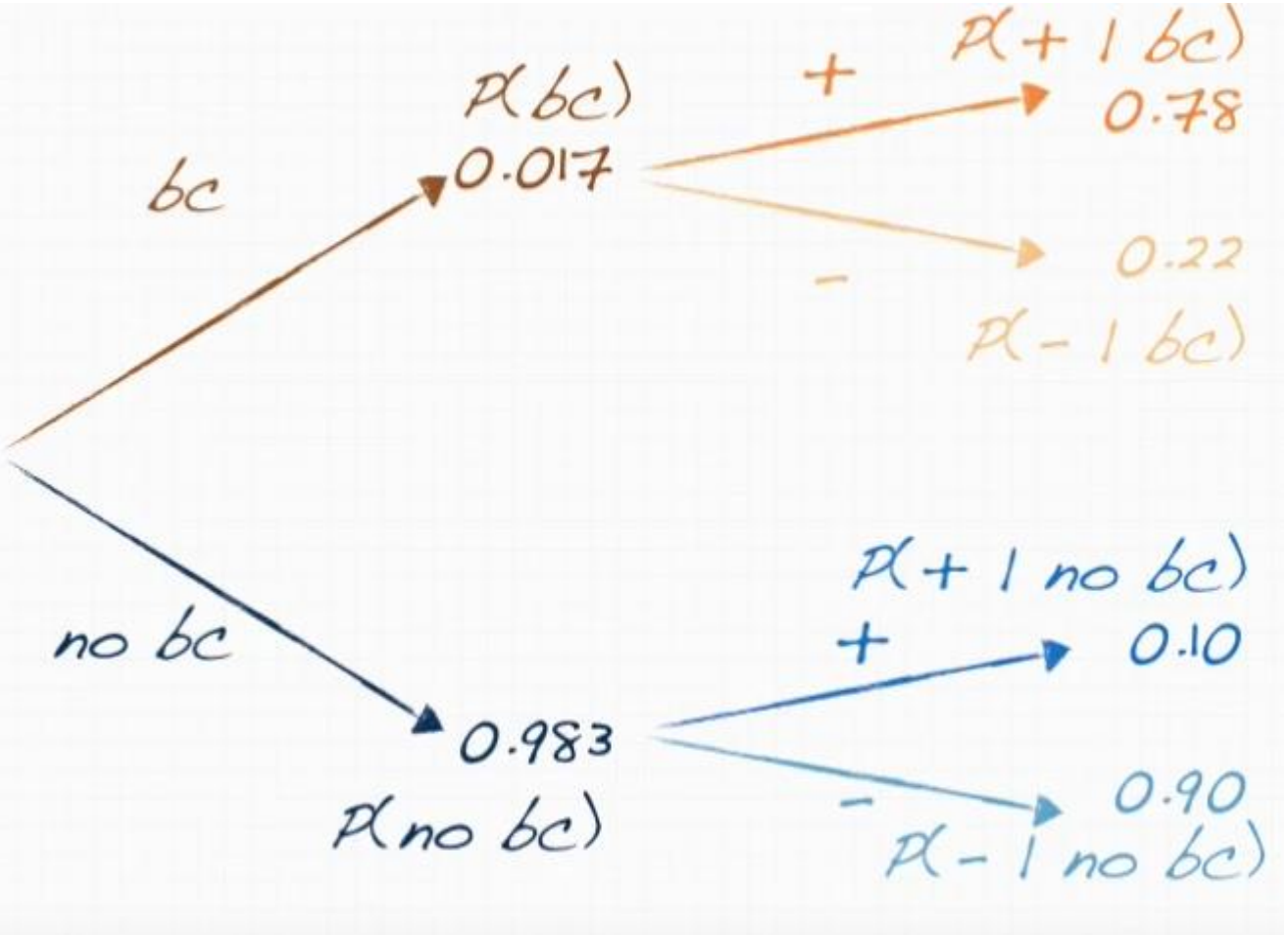
When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?



$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0133}{0.0133 + 0.0983} = 0.12$$

Note: Tree diagrams are useful for inverting probabilities: we are given P(+|C) and asked for P(C|+).

When a patient goes through breast cancer screening there are two competing claims: patient has cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient has cancer? $P(bc \mid +) = ?$



$P(bc \text{ and } +)$

$0.017 \times 0.78 = 0.01326$

$$P(bc \mid +) = \frac{P(bc \text{ and } +)}{P(+)}$$

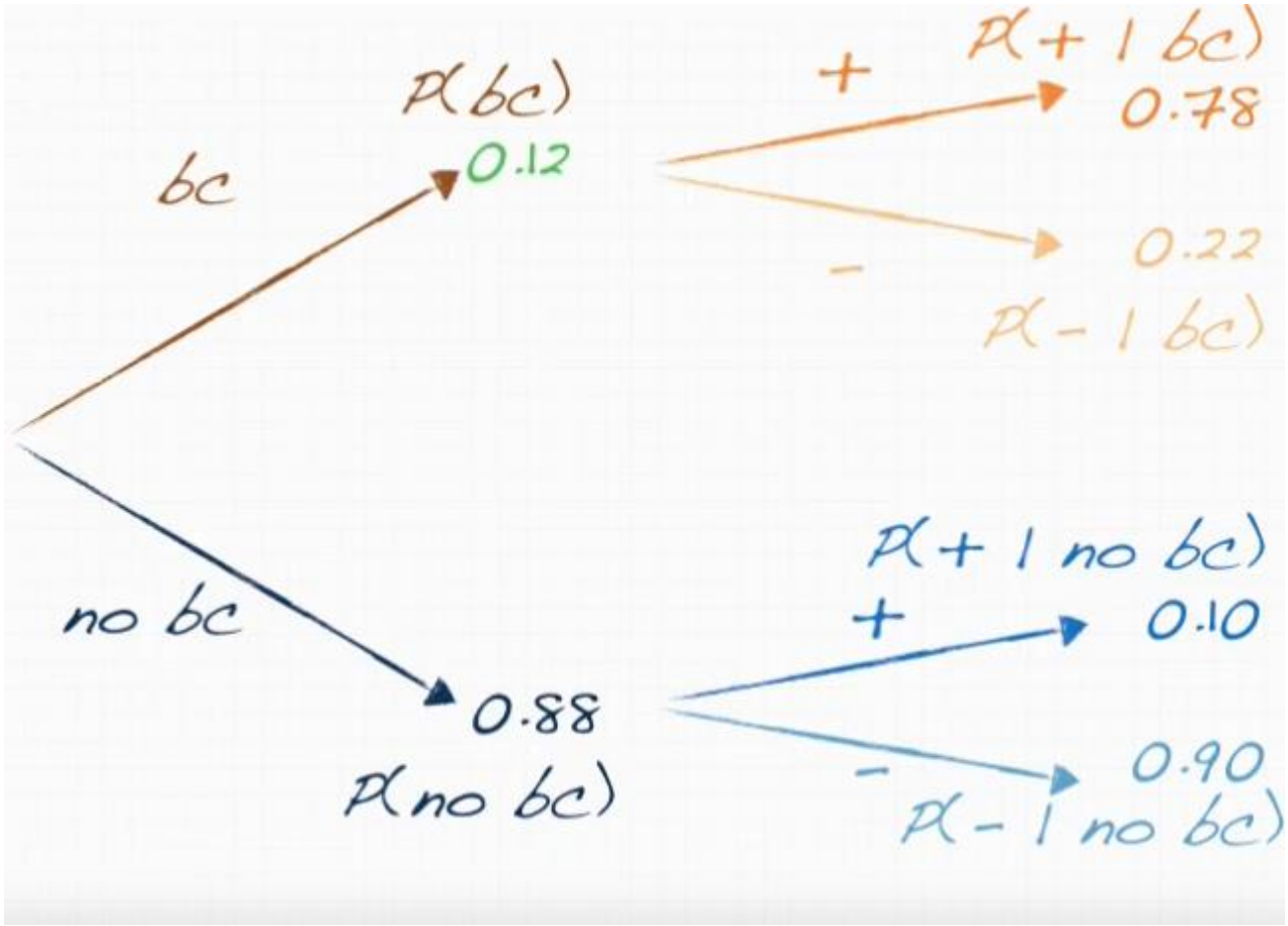
$$= \frac{0.01326}{0.01326 + 0.0983} \approx 0.12$$

posterior

$0.983 \times 0.10 = 0.0983$

$P(\text{no } bc \text{ and } +)$

Since a positive mammogram doesn't necessarily mean that the patient actually has breast cancer, the doctor might decide to re-test the patient. What is the probability of having breast cancer if this second mammogram also yields a positive result?



P(bc and +)

$0.12 \times 0.78 = 0.0936$

$0.88 \times 0.10 = 0.088$

P(no bc and +)

P(bc | +) = 0.0936 / (0.0936+0.088)

≈ 0.52

Practice

Suppose a woman who gets tested once and obtains a positive result wants to get tested again. In the second test, what should we assume to be the probability of this specific woman having cancer?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Practice

Suppose a woman who gets tested once and obtains a positive result wants to get tested again. In the second test, what should we assume to be the probability of this specific woman having cancer?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Practice

What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52

Practice

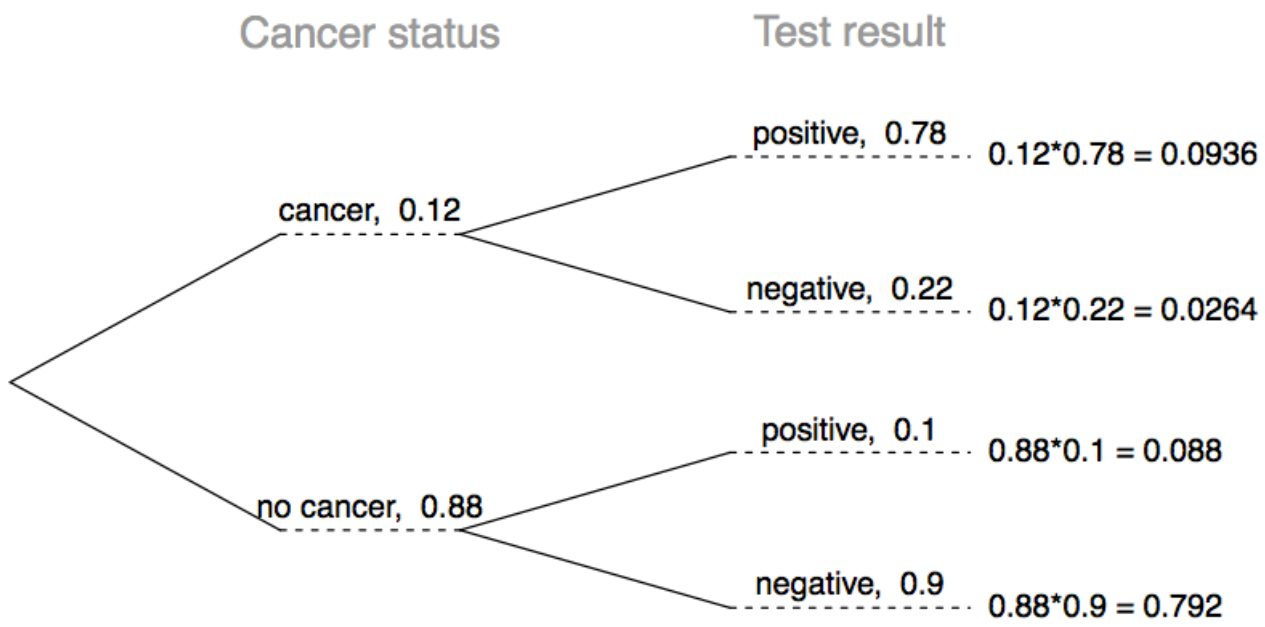
What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52



Practice

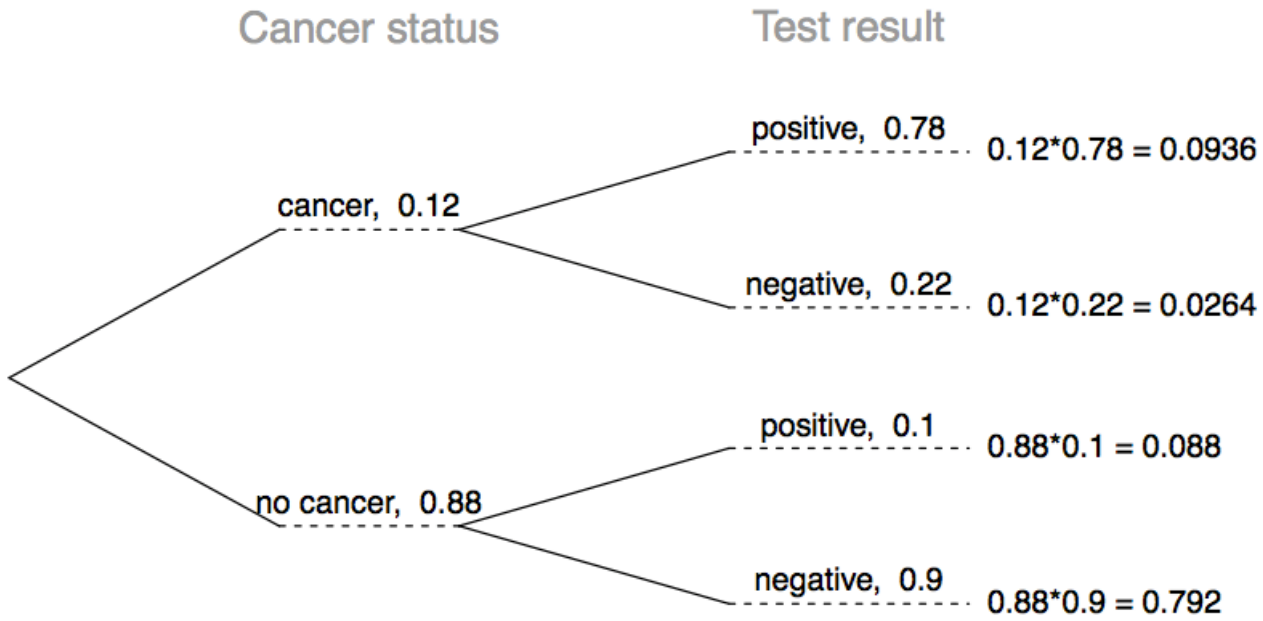
What is the probability that this woman has cancer if this second mammogram also yielded a positive result?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52



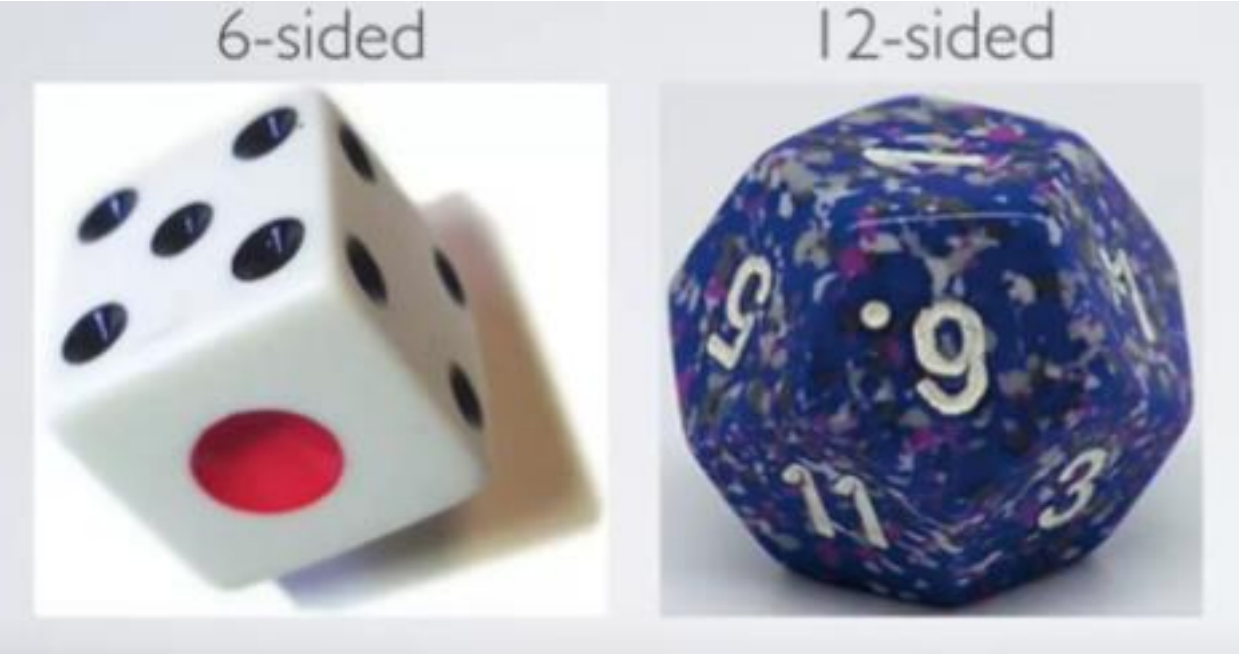
$$P(C|+) = \frac{P(C \text{ and } +)}{P(+)} = \frac{0.0936}{0.0936 + 0.088} = 0.52$$

Bayes Inference – recap

- setting a prior
- collecting data
- obtaining a posterior
- updating the prior with the previous posterior

Bayesian Inference example 2

Set-up



Problem: Player holds one die in each hand. ***Need to figure out which hand holds which die.***

Note: This is not a guessing game. Before you make a decision, you will be able to collect data, and ask the player to roll the die and confirm to you whether the **outcome of the die roll is >= 4**

Probabilities

What is the probability of rolling ≥4 with a 6-sided die?

$S = \{1,2,3,4,5,6\}$

$P(\geq 4) = 3/6 = 1/2 = 0.5$

What is the probability of rolling ≥4 with a 12-sided die?

$S = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$P(\geq 4) = 9/12 = 3/4 = 0.75$

"good die"

Say you're playing a game where the goal is to roll ≥ 4. If you could get your pick, which die would you prefer to play this game with?

(a)

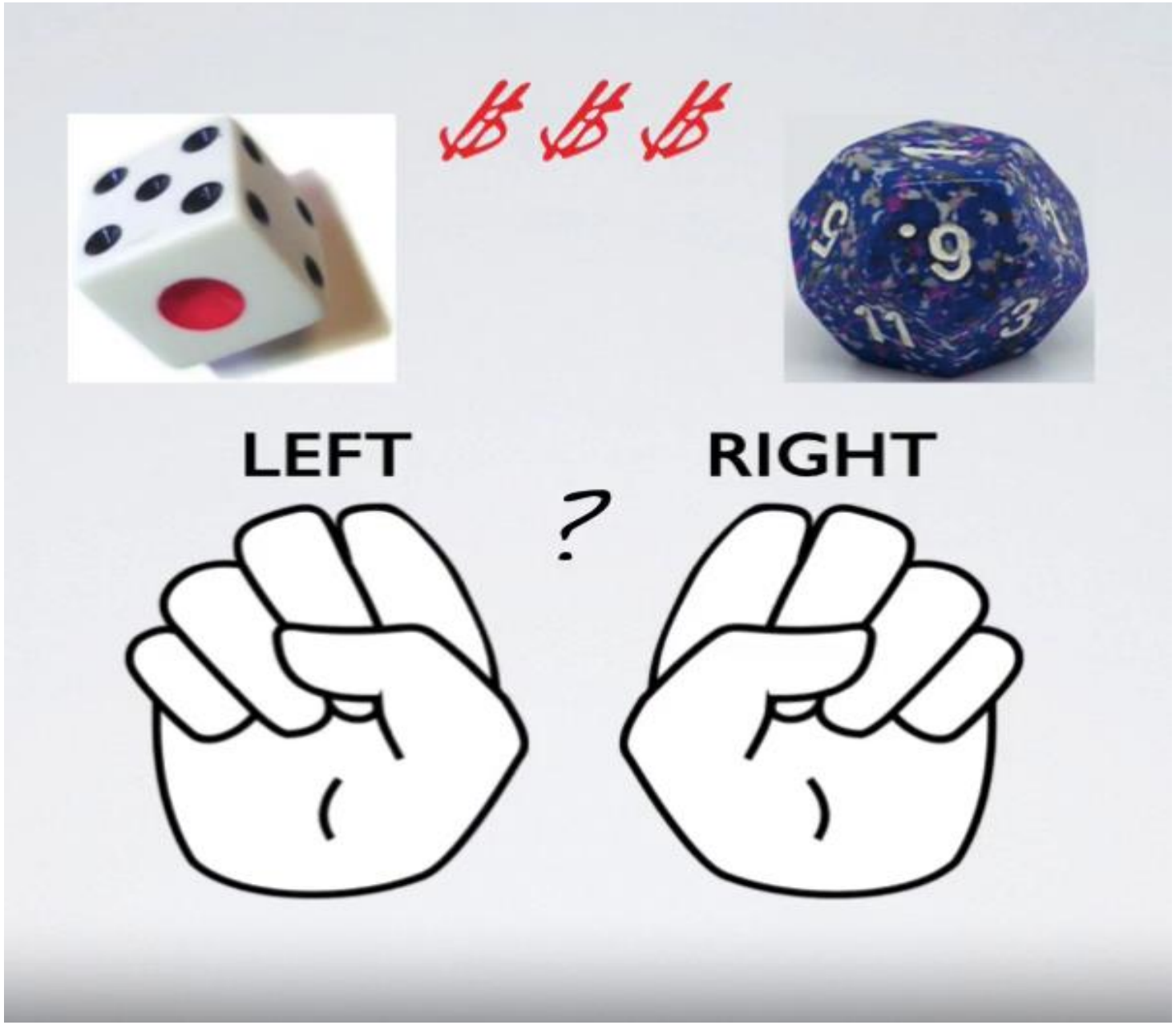
$P(\geq 4) = 0.5$

good die

(b)

$P(\geq 4) = 0.75$

Rules



Hypotheses and decisions

		Truth	
		Right good, Left bad	Right bad, Left good
Decision	pick Right	You win the game!	You lose :(
	pick Left	You lose :(	You win the game!

cost of losing

certainty from more data

before you collect data

Before we collect any data, you have no idea if I am holding the good die (12-sided) on the right hand or the left hand. Then, what are the probabilities associated with the following hypotheses?

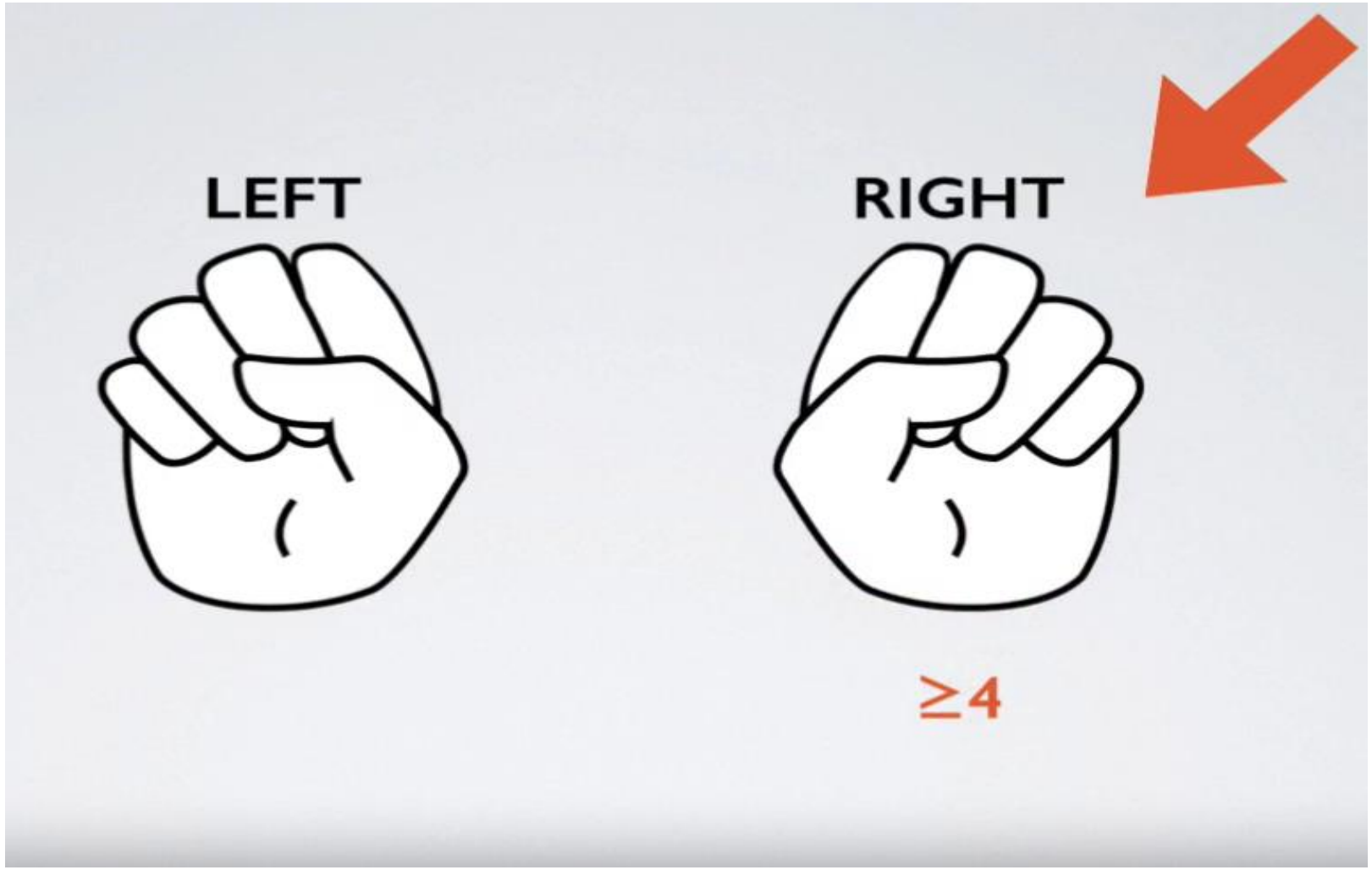
H_1: good die on the Right (bad die on the Left)

H_2: good die on the Left (bad die on the Right)

	P(H_1: good die on the Right)	P(H_2: good die on the Left)
(a)	0.33	0.67
(b)	0.5	0.5
(c)	0	1
(d)	0.25	0.75

→ prior

data collection: round # 1



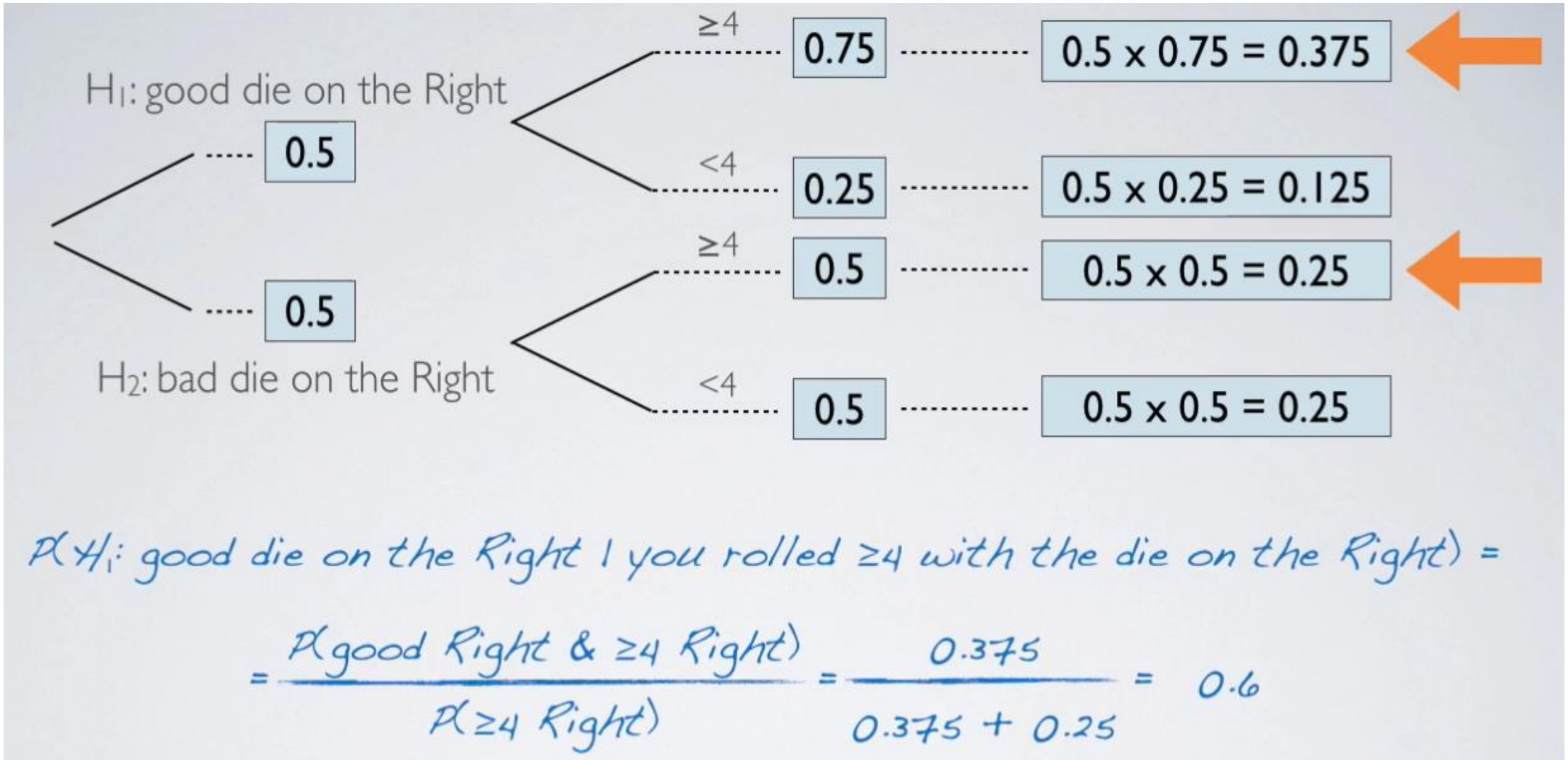
after you see the data

You chose the right hand, and you won (rolled a number ≥4). Having observed this data point how, if at all, do the probabilities you assign to the same set of hypotheses change?

H_1: good die on the Right (bad die on the Left)

H_2: good die on the Left (bad die on the Right)

	$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
(a)	0.5	0.5
(b)	more than 0.5	less than 0.5
(c)	less than 0.5	more than 0.5



$P(H_1$: good die on the Right | you rolled ≥4 with the die on the Right) =

$$= \frac{P(\text{good Right \& } \geq 4 \text{ Right})}{P(\geq 4 \text{ Right})} = \frac{0.375}{0.375 + 0.25} = 0.6$$

Posterior probability

- The probability we just calculated is also called the posterior probability.

  $P(H_1$: good die on the Right | you rolled ≥ 4 with the die on the Right)

- Posterior probability is generally defined as P(hypothesis | data).
- It tells us the probability of a hypothesis we set forth, given the data we just observed.
- It depends on both the prior probability we set and the observed data.

Updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data.
- In the next iteration (roll) we get to take advantage of what we learned from the data.
- In other words, we update our prior with our posterior probability from the previous iteration.

updated:

$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
0.6	0.4

recap

- Take advantage of prior information, like a previously study or a physical model
- Naturally integrate data as you collect it, and update your priors
- Base decisions on the posterior probability:

P(hypothesis is true | observed data)

- A good prior helps, a bad prior hurts, but the prior matters less the more data that you have.

Bayes' Theorem: Inverting Probabilities

BAYES' THEOREM: INVERTING PROBABILITIES

Consider the following conditional probability for variable 1 and variable 2:

$$P(\text{outcome } A_1 \text{ of variable 1} \mid \text{outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable.

Bayes' Theorem – contd.

Bayes' Theorem is a generalization of what we have done using tree diagrams. The numerator identifies the probability of getting both A1 and B. The denominator is the marginal probability of getting B. This bottom component of the fraction appears long and complicated since we have to add up probabilities from all of the different ways to get B.

To apply Bayes' Theorem correctly, there are two preparatory steps:

(1) First identify the marginal probabilities of each possible outcome of the first variable: $P(A_1)$, $P(A_2)$, ..., $P(A_k)$.

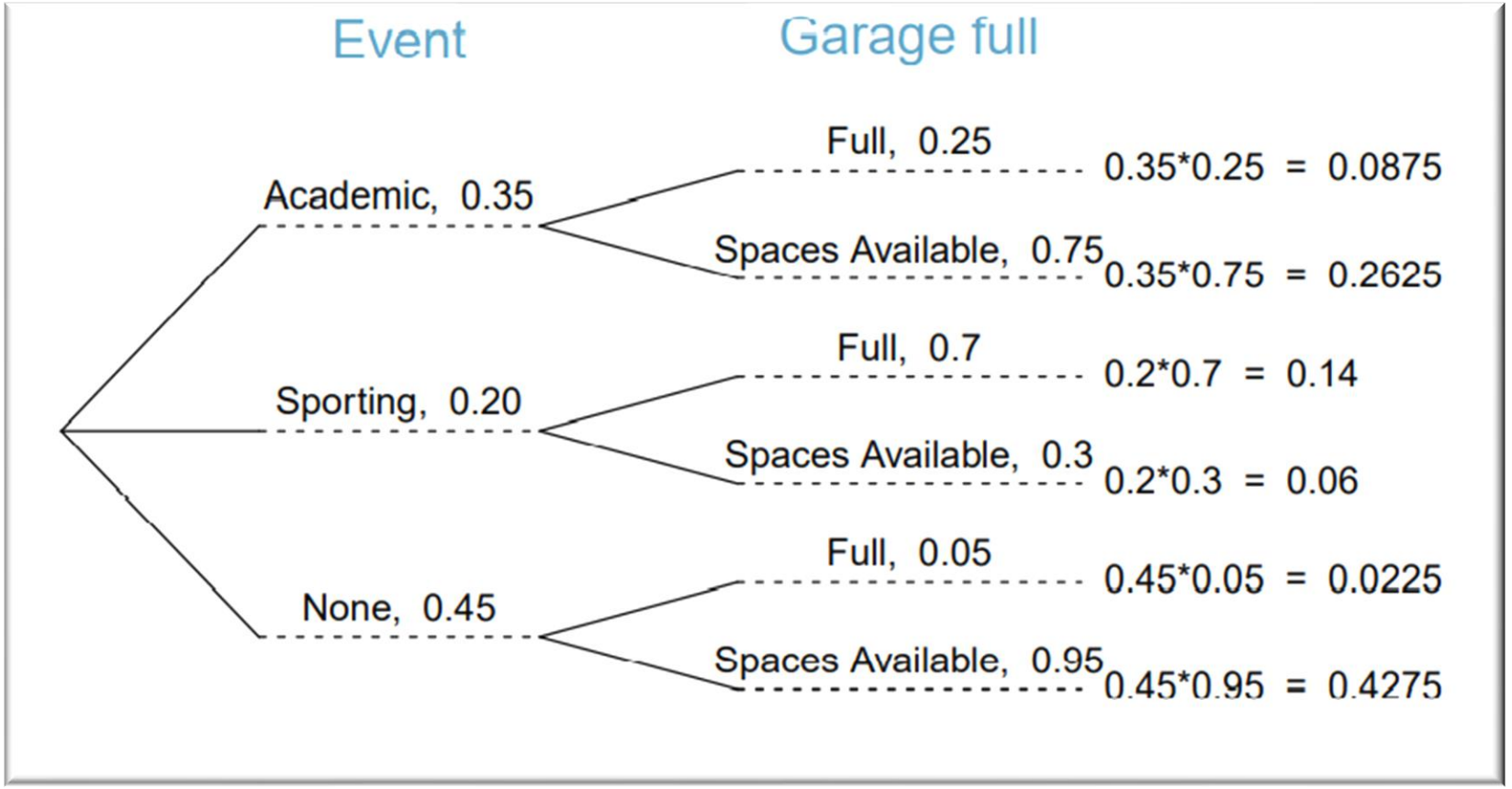
(2) Then identify the probability of the outcome B, conditioned on each possible scenario for the first variable: $P(B|A_1)$, $P(B|A_2)$, ..., $P(B|A_k)$.

Once each of these probabilities are identified, they can be applied directly within the formula. *Bayes' Theorem tends to be a good option when there are so many scenarios that drawing a tree diagram would be complex.*

Practice Problem

Jose visits campus every Thursday evening. However, some days the parking garage is full, often due to college events. There are **academic events** on 35% of evenings, **sporting events** on 20% of evenings, and **no events** on 45% of evenings. When there is an academic event, the garage fills up about 25% of the time, and it fills up 70% of evenings with sporting events. On evenings when there are no events, it only fills up about 5% of the time. If Jose comes to campus and finds the garage full, what is the probability that there is a sporting event? **Solve the problem using (a) a tree diagram, (b) Bayes' Theorem.**

Solution(a) Using Tree Diagram



The tree diagram, with three primary branches, is shown above. Next, we identify two probabilities from the tree diagram. (1) The probability that there is a sporting event and the garage is full: 0.14. (2) The probability the garage is full: 0.0875 + 0.14 + 0.0225 = 0.25. Then the solution is the ratio of these probabilities: 0.14 / 0.25 = 0.56. If the garage is full, there is a 56% probability that there is a sporting event.

Solution(b) Using Bayes' Theorem

The outcome of interest is whether there is a sporting event (call this A_1), and the condition is that the lot is full (B). Let A_2 represent an academic event and A_3 represent there being no event on campus. Then the given probabilities can be written as

$$P(A_1) = 0.2 \qquad P(A_2) = 0.35 \qquad P(A_3) = 0.45$$
$$P(B|A_1) = 0.7 \qquad P(B|A_2) = 0.25 \qquad P(B|A_3) = 0.05$$

Bayes' Theorem can be used to compute the probability of a sporting event (A_1) under the condition that the parking lot is full (B):

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.7)(0.2)}{(0.7)(0.2) + (0.25)(0.35) + (0.05)(0.45)} \\ &= 0.56 \end{aligned}$$

Based on the information that the garage is full, there is a 56% probability that a sporting event is being held on campus that evening.

Random Variables

Random variables

It’s often useful to model a process using what’s called a random variable. Such a model allows us to apply a mathematical framework and statistical principles for better understanding and predicting **outcomes in the real world.**

A *random variable* is a numeric quantity whose value depends on the outcome of an underlying random event

- We use a capital letter, like **X**, to denote a random variable
- The values of a random variable are denoted with a lowercase letter, in this case **x**

Random variables - Interpretation

Informally, a random variable assigns numbers to outcomes in the sample space. *So, instead of focusing on the outcomes themselves, we highlight a specific characteristic of the outcomes.*

Example: Suppose we toss a coin twice and record the sequence of heads (h) and tails (t). The sample space for this random experiment is given by:

$$S = \{hh, ht, th, tt\}.$$

Suppose we are only interested in tosses that result in heads. We can define a random variable **X** that tracks the number of heads obtained in an outcome.

So, if outcome hh is obtained, then **X** will equal 2. Formally, we denote this as follows:

$$X : S \to \mathbb{R}$$
$$s \mapsto \text{number of } h\text{'s in } s$$

Since there are only four outcomes in S , we can list the value of X for each outcome individually:

$$\begin{array}{lcl} \text{inputs: } S & \overset{\text{function: } X}{\longrightarrow} & \text{outputs: } \mathbb{R} \\ hh & \overset{X}{\mapsto} & 2 \\ th & \overset{X}{\mapsto} & 1 \\ ht & \overset{X}{\mapsto} & 1 \\ tt & \overset{X}{\mapsto} & 0 \end{array}$$

We can also write the above as follows:

$$X(hh) = 2, \quad X(ht) = X(th) = 1, \quad X(tt) = 0.$$

The advantage to defining the random variable X in this context is that *the two outcomes ht and th are both assigned a value of 1* , meaning we are not focused on the actual sequence of heads and tails that resulted in obtaining one head.

In the above example, note that the random variable we defined only equals one of three possible values: $0,1,2$. This is an example of what we call a **discrete random variable**.

Example:

I toss a coin five times. This is a random experiment and the sample space can be written as

$$S = \{TTTTT, TTTTH, \ldots, HHHHH\}.$$

Note that here the sample space S has $2^5 = 32$ elements. Suppose that in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed heads. The value of X will be one of $0, 1, 2, 3, 4$ or 5 depending on the outcome of the random experiment.

- In essence, a random variable is a real-valued function that assigns a numerical value to each possible outcome of the random experiment.
- For example, the random variable X defined above assigns the value 0 to the outcome **TTTTT**, the value 2 to the outcome THTHT, and so on.
- The random variable X is a function from the sample space **S = { TTTTT, TTTTH, ⋯ , HHHHH }** to the real numbers.
- For this particular random variable, the values are always integers between 0 and 5.

Random Variables:

A random variable X is a function from the sample space to the real numbers.

$$X : S \to \mathbb{R}$$

The range of a random variable X, shown by Range(X) or R_X, is the set of possible values of X.

In the above example, Range(X) = { 0, 1, 2, 3, 4, 5 }.

Example

Find the range for each of the following random variables.

1. I toss a coin 100 times. Let X be the number of heads I observe.
2. I toss a coin until the first heads appears. Let Y be the total number of coin tosses.
3. The random variable T is defined as the time (in hours) from now until the next earthquake occurs in a certain city.

Solution

1. The random variable X can take any integer from 0 to 100, so $R_X = \{0, 1, 2, \ldots, 100\}$.
2. The random variable Y can take any positive integer, so $R_Y = \{1, 2, 3, \ldots\} = \mathbb{N}$.
3. The random variable T can in theory get any nonnegative real number, so $R_T = [0, \infty)$.

Discrete Random Variable - Definition

A random variable **X** is said to be **discrete** if its range consists of a **finite** or **countable** number of values.

Note: A set **A** is countable if either:

- A is a finite set such as **{1, 2, 3, 4}**, or
- it can be put in one-to-one correspondence with natural numbers (in this case the set is said to be countably infinite)

In the previous slide, the random variables X and Y are discrete, while the random variable T is not discrete

Example: based on tossing a coin repeatedly

- No. of H in 1st 5 tosses: {0, 1, 2, ..., 5}
- No. of T before first H: {0, 1, 2, ...}

Probability Mass Function(pmf) - Definition

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \ldots\}$ (finite or countably infinite). The function

$$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3, \ldots,$$

is called the *probability mass function (PMF)* of X.

Note: The PMF is a probability measure that gives us probabilities of the possible values for a random variable. The subscript X here indicates that this is the PMF of the random variable X. Thus, for example, PX(1) shows the probability that X=1.

Probability Mass Function – cont.

If X is a discrete random variable then its range R_X is a countable set, so, we can list the elements in R_X. In other words, we can write

$$R_X = \{x_1, x_2, x_3, \ldots\}.$$

The event $A = \{X = x_k\}$ is defined as the set of outcomes s in the sample space S for which the corresponding value of X is equal to x_k. In particular,

$$A = \{s \in S | X(s) = x_k\}.$$

The probabilities of events $\{X = x_k\}$ are formally shown by the **probability mass function (pmf)** of X.

Properties of PMF

Consider a discrete random variable X with Range$(X) = R_X$. Note that by definition the PMF is a probability measure, so it satisfies all properties of a probability measure. In particular, we have

- $0 \leq P_X(x) \leq 1$ for all x, and
- $\sum_{x \in R_X} P_X(x) = 1$.

Also note that for any set $A \subset R_X$, we can find the probability that $X \in A$ using the PMF

$$P(X \in A) = \sum_{x \in A} P_X(x).$$

Example

I toss a fair coin twice, and let X be defined as the number of heads I observe. Find the range of X, R_X, as well as its probability mass function P_X.

Here our sample space is **S = {HH, HT, TH, TT}.**

The number of heads will be **0, 1** or **2**. Thus $R_X = \{0, 1, 2\}.$

Since this is a finite (and thus a countable) set, the random variable X is a discrete random variable. Next, we need to find PMF of X. The PMF is defined as

$$P_X(k) = P(X = k) \text{ for } k = 0, 1, 2.$$

We have

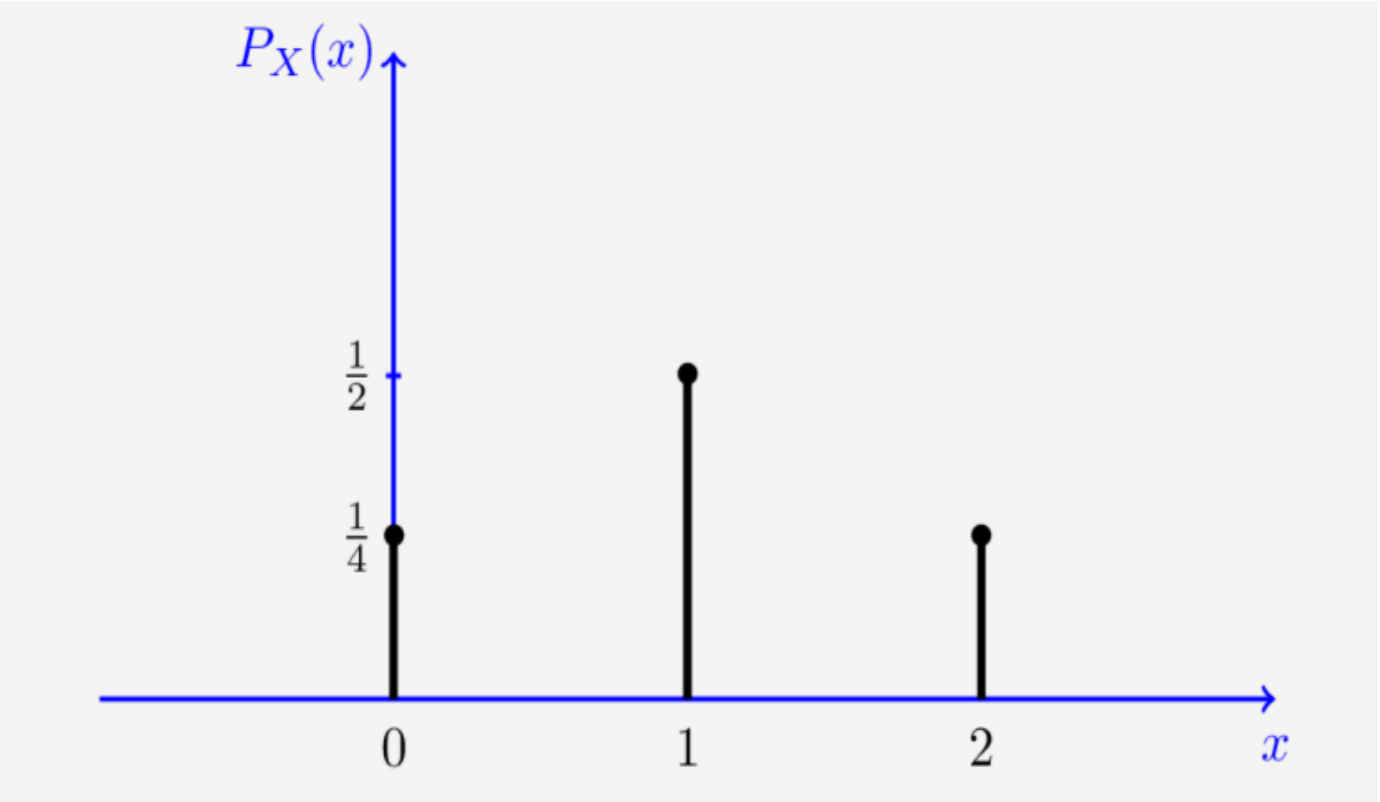
$$P_X(0) = P(X = 0) = P(TT) = \frac{1}{4},$$

$$P_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

$$P_X(2) = P(X = 2) = P(HH) = \frac{1}{4}.$$

Example – contd.

To better visualize the PMF, we can plot it. The Figure shows the PMF of the above random variable X. As we see, the random variable can take three possible values 0,1 and 2.



Example

I have an unfair coin for which $P(H) = p$, where $0 < p < 1$. I toss the coin repeatedly until I observe a heads for the first time. Let Y be the total number of coin tosses. Find the distribution of Y.

First, we note that the random variable Y can potentially take any positive integer, so we have $R_Y = \mathbb{N} = \{1, 2, 3, \dots\}$. To find the distribution of Y, we need to find $P_Y(k) = P(Y = k)$ for $k = 1, 2, 3, \dots$. We have

$$P_Y(1) = P(Y = 1) = P(H) = p,$$

$$P_Y(2) = P(Y = 2) = P(TH) = (1-p)p,$$

$$P_Y(3) = P(Y = 3) = P(TTH) = (1-p)^2 p,$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$P_Y(k) = P(Y = k) = P(TT\dots TH) = (1-p)^{k-1} p.$$

Thus, we can write the PMF of Y in the following way

$$P_Y(y) = \begin{cases} (1-p)^{y-1} p & \text{for } y = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

Cumulative Distribution Function

The PMF is one way to describe the distribution of a discrete random variable. However, *PMF cannot be defined for continuous random variables*. The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. The advantage of the CDF is that it can be defined for **any kind of random variable** (discrete, continuous, and mixed).

Definition:

The cumulative distribution function (CDF) of random variable X is defined as

$$F_X(x) = P(X \leq x), \text{ for all } x \in \mathbb{R}.$$

Note: the subscript **X** indicates that this is the CDF of the random variable X. Also, note that the CDF is defined for all **x∈R**.

Expectation

If you have a collection of numbers $a_1, a_2, \ldots, a_N$, their average is a single number that describes the whole collection. Now, consider a random variable X. We would like to define its average, or as it is called in probability, its **expected value** or **mean**. The expected value is defined as the weighted average of the values in the range.

Definition:

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \ldots\}$ (finite or countably infinite). The *expected* value of X, denoted by EX is defined as

$$EX = \sum_{x_k \in R_X} x_k P(X = x_k) = \sum_{x_k \in R_X} x_k P_X(x_k).$$

To understand the concept behind EX, consider a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots\}$. This random variable is a result of random experiment. Suppose that we repeat this experiment a very large number of times N, and that the trials are independent. Let N_1 be the number of times we observe x_1, N_2 be the number of times we observe x_2, ..., N_k be the number of times we observe x_k, and so on. Since $P(X = x_k) = P_X(x_k)$, we expect that

$$
\begin{aligned}
P_X(x_1) &\approx \frac{N_1}{N}, \\
P_X(x_2) &\approx \frac{N_2}{N}, \\
&\cdot \quad \cdot \quad \cdot \\
P_X(x_k) &\approx \frac{N_k}{N}, \\
&\cdot \quad \cdot \quad \cdot
\end{aligned}
$$

In other words, we have $N_k \approx NP_X(x_k)$. Now, if we take the average of the observed values of X, we obtain

$$\begin{aligned}
\text{Average} &= \frac{N_1x_1+N_2x_2+N_3x_3+\ldots}{N} \\
&\approx \frac{x_1NP_X(x_1)+x_2NP_X(x_2)+x_3NP_X(x_3)+\ldots}{N} \\
&= x_1P_X(x_1)+x_2P_X(x_2)+x_3P_X(x_3)+\ldots \\
&= EX.
\end{aligned}$$

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the *expected value* (mean), and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Example: Expected value of a discrete random variable

In a game of cards you win $1 if you draw a heart, $5 if you draw an ace (including the ace of hearts), $10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

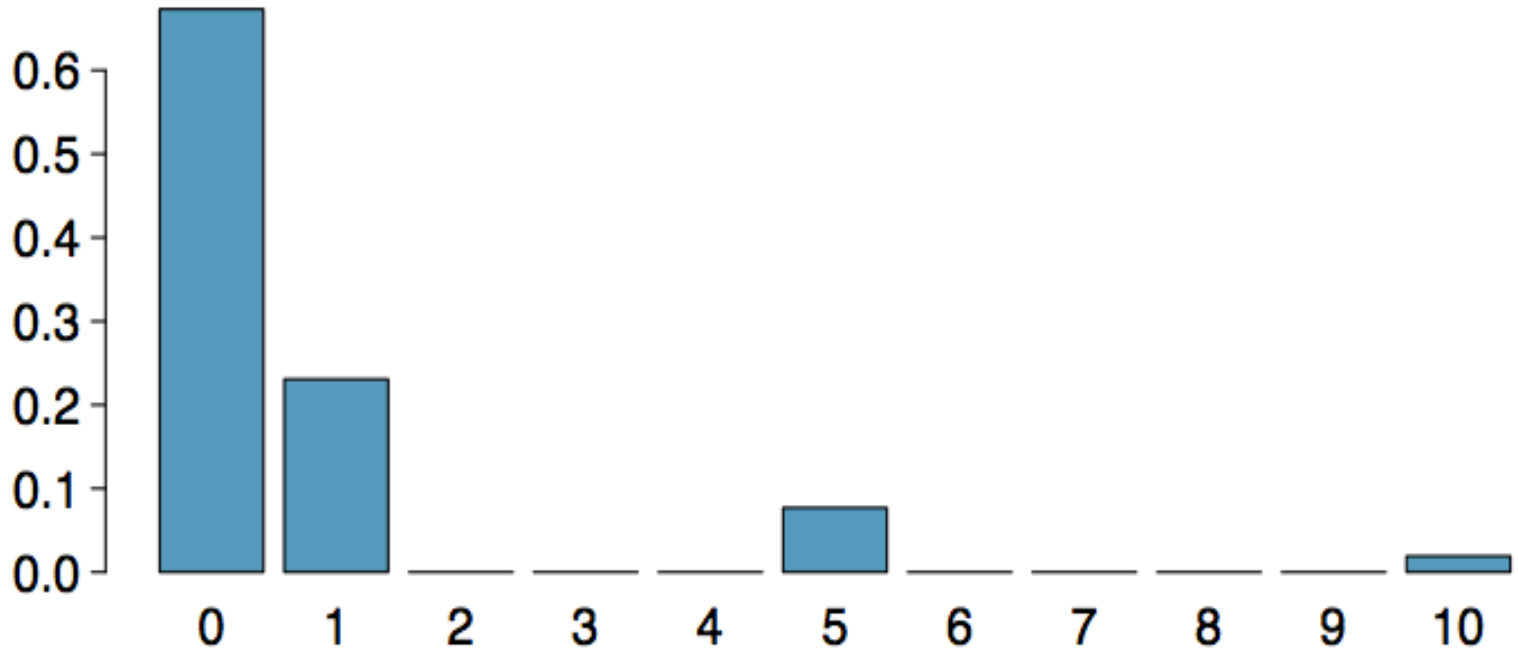
Expected value of a discrete random variable

In a game of cards you win $1 if you draw a heart, $5 if you draw an ace (including the ace of hearts), $10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:



Variability

We are also often interested in the variability in the values of a random variable.

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246} = 1.85$

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

  where a and b are some fixed numbers.

- The average value of a linear combination of random variables is given by

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

$$
\begin{aligned}
E(S+S+S+S+S+C+C+C+C) &= 5 \times E(S) + 4 \times E(C) \\
&= 5 \times 10 + 4 \times 15 \\
&= 50 + 60 \\
&= 110\ min
\end{aligned}
$$

Linear Combination

- The variability of a linear combination of two independent random variables is calculated as:

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

Note: If the random variables are not independent, the variance calculation gets a little more complicated and is beyond the scope of this course.

Linear Combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and chemistry homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

Linear Combination (can you spot the error?)

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and chemistry homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

V(5 x S + 4 x C)

$$
\begin{aligned}
& V(S+S+S+S+S+C+C+C+C) \\
&= V(S)+V(S)+V(S)+V(S)+V(S)+V(C)+V(C)+V(C)+V \\
&= 5 \times V(S)+4 \times V(C) \\
&= 5 \times 1.5^2+4 \times 2^2 \\
&= 27.25
\end{aligned}
$$