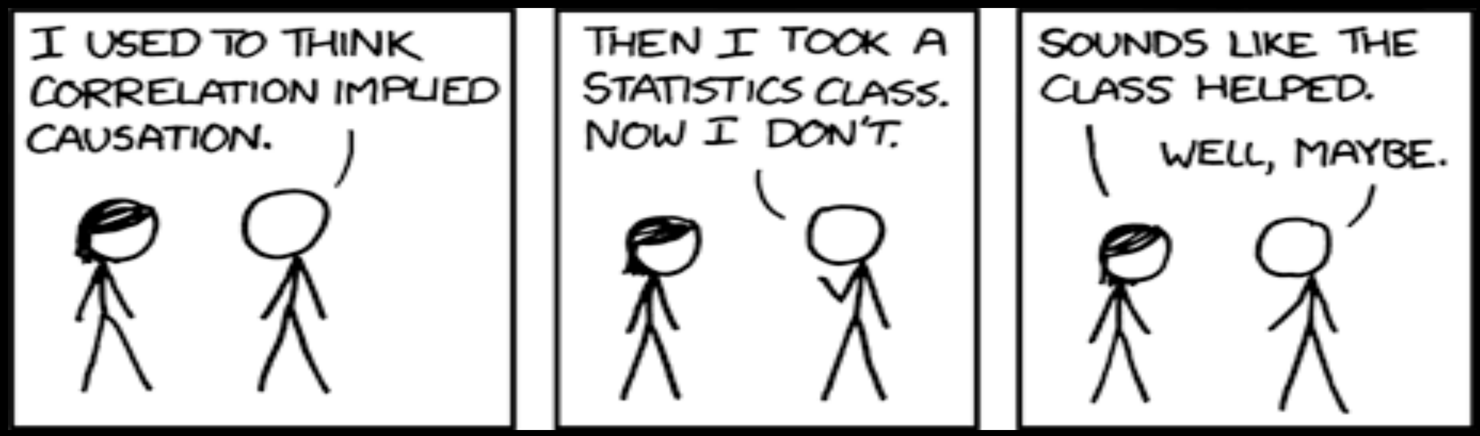
Welcome to Week #4!

Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

Definitions, basic concepts, R practice

Correlation is not causation

- Correlation is not causation!



http:// xkcd.com/ 552/

- Observational studies alone cannot prove causation; only well designed experiments can prove causation.

Observational & Experimental Studies: Summary

1. Terminology:

sample vs. population
observational vs. experimental studies
explanatory vs. response variables
confounding factors
blocking factors
placebo, placebo effect
blinding, double blinding
association vs. casually connected

2. Table Proportions

e.g. percentage of healthy patients after receiving placebo vs. treatment

3. Sampling Methods (section 1.4)

Is the survey given out randomly? How are participants selected?

Intro to Probability Theory: A bunch of definitions & problems

(lots of definitions & equations, followed by some playing of online games)

Probability: Practically

What is the probability of event #1 **or** event #2 occurring?

- $P(E_1 \text{ or } E_2)$ - General Addition Rule

What is the probability of event #2 **and** event #2 occurring?

- $P(E_1 \text{ and } E_2)$ - General Multiplication Rule

What is the probability of event #1 **given** event #2 (if event #1 depends on event #2)?

- $P(E_1 \mid E_2)$ - Conditional probability - marginal & joint probabilities; tree diagrams; Bayes' Theorem

How do these relate $P(E_1)$ and $P(E_2)$?

2 rules of probability

$0 \le P(E) \le 1$ (or $0\% \le P(E) \le 100\%$)

0 implies the event is impossible and
1 implies the event is certain

Also, P(E) + P(not E) = 1 and so

P(not E) = 1 - P(E)

e.g. P(not getting a Queen in a deck of 52 cards)
=1 – P(getting a Queen in a deck of 52 cards)
= 1 – 4/52 = 48/52

Disjoint and non-disjoint outcomes

Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

Non-disjoint outcomes: Can happen at the same time.

- A student can get an A in Stats and A in Programming in the same semester.

Recap

General addition rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Note: For disjoint/mutually exclusive events, $P(A \text{ and } B) = 0$, so the above formula simplifies to $P(A \text{ or } B) = P(A) + P(B)$

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

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

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

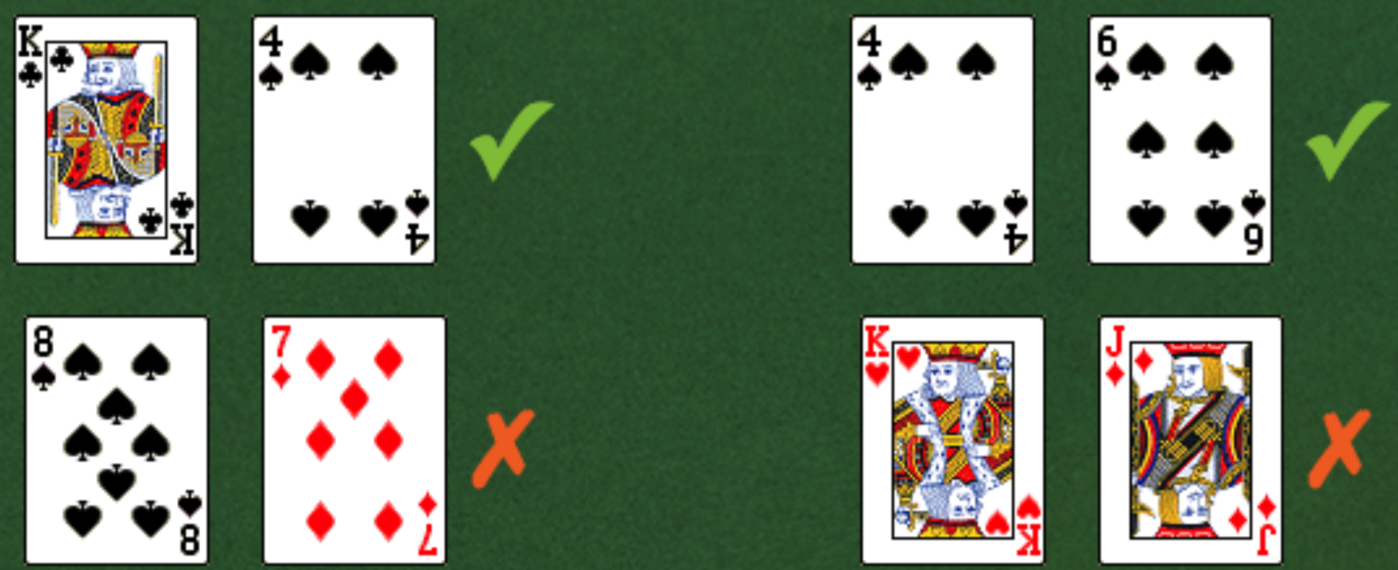
Bank: $171
Level 3

Beat the Odds

Training

You draw 2 cards from a deck. What's the probability that they are both black?

EXAMPLES



Go to the Lab »

YOUR ANSWER

$p =$

Submit

« Back to menu

Skip this problem

Bank: $174
Level 3

Beat the Odds

Training

You flip 3 coins. What's the probability that none are tails?

EXAMPLES

Go to the Lab »

YOUR ANSWER

$p =$

Submit

Skip this problem

« Back to menu

Bank: $219
Level 4

Beat the Odds

Training

You roll 3 dice. What's the probability that at least one roll equals 3?

EXAMPLES

Go to the Lab »

YOUR ANSWER

$p =$

Submit

Skip this problem

« Back to menu

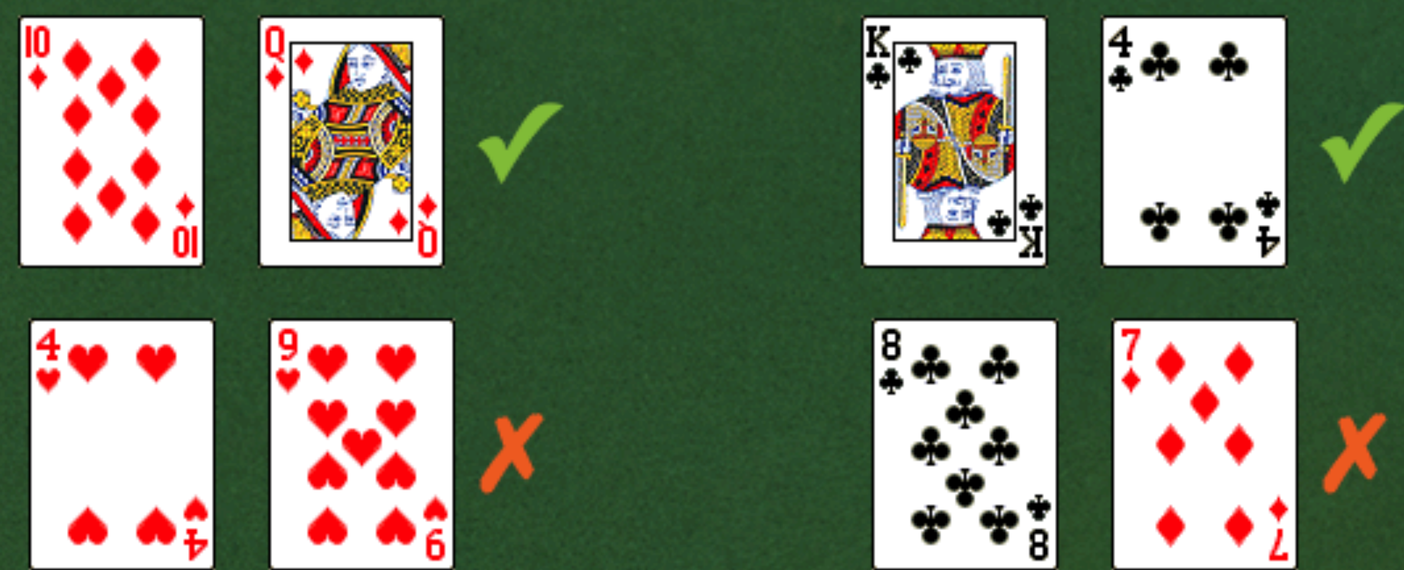
Bank: $260
Level 4

Beat the Odds

Training

You draw 2 cards from a deck. What's the probability that exactly one is a face card (Jack, Queen, or King)?

EXAMPLES



Go to the Lab »

YOUR ANSWER

$p =$

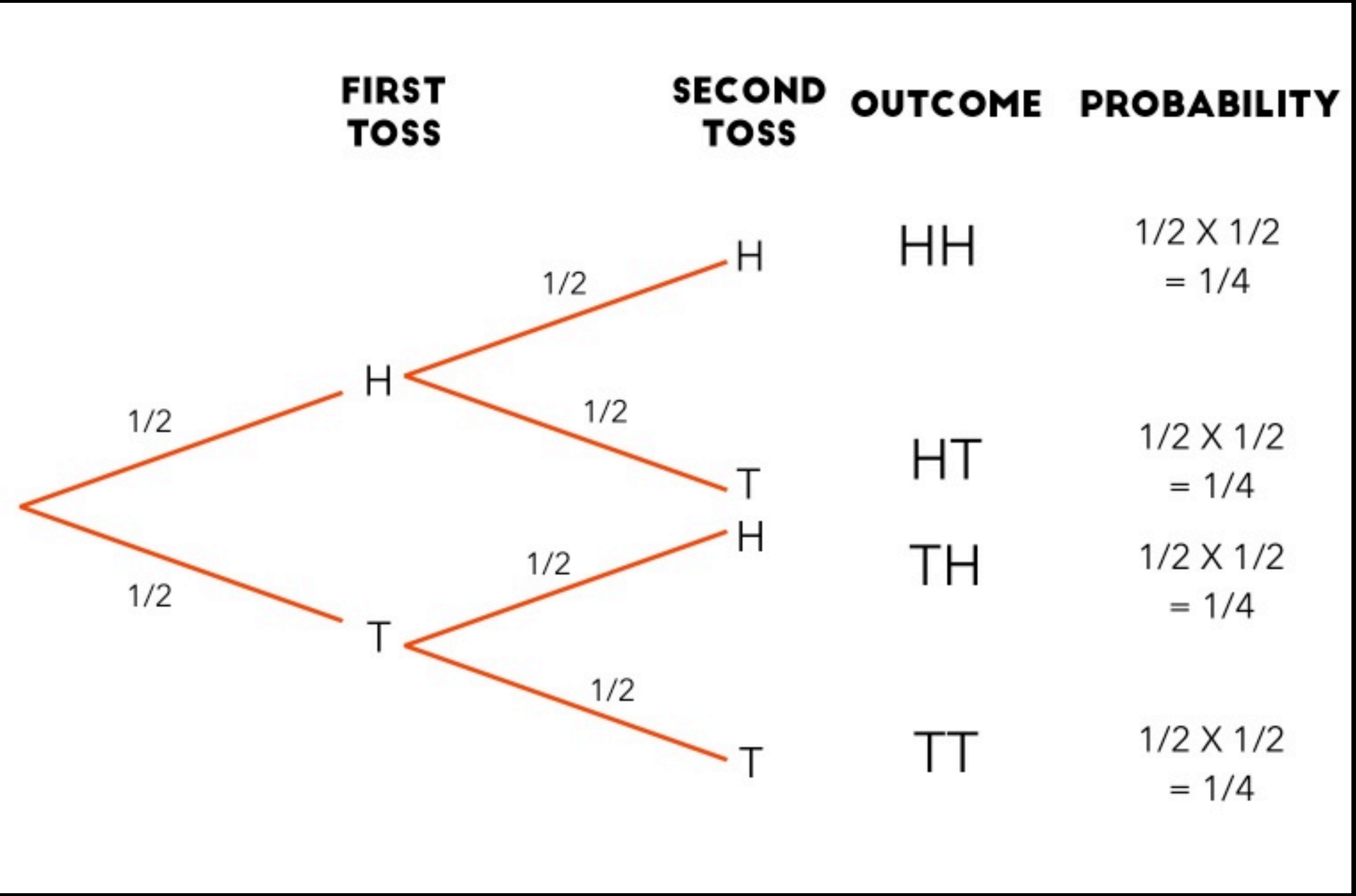
Submit

« Back to menu

Skip this problem

Beat the Odds Game:
http://d3tt741pwxqwm0.cloudfront.net/WGBH/mgbh/mgbh_int_beatodds/index.html

Tree Diagrams: inverting probabilities



Application activity: inverting probabilities

- A common epidemiological model for the spread of diseases is the SIR model, where the population is partitioned into three groups: Susceptible, Infected, and Recovered. This is a reasonable model for diseases like chickenpox where a single infection usually provides immunity to subsequent infections. Sometimes these diseases can also be difficult to detect.

- Imagine a population in the midst of an epidemic where 60% of the population is considered susceptible, 10% is infected, and 30% is recovered. The only test for the disease is accurate 95% of the time for susceptible individuals, 99% for infected individuals, but 65% for recovered individuals. (Note: In this case accurate means returning a negative result for susceptible and recovered individuals and a positive result for infected individuals).

- Draw a probability tree to reflect the information given above. If the individual has tested positive, what is the probability that they are actually infected?

Application activity: inverting probabilities

- Review of conditional probability relation/general multiplication law

$$P(A \mid B) = P(A \,\&\, B)/P(B)$$

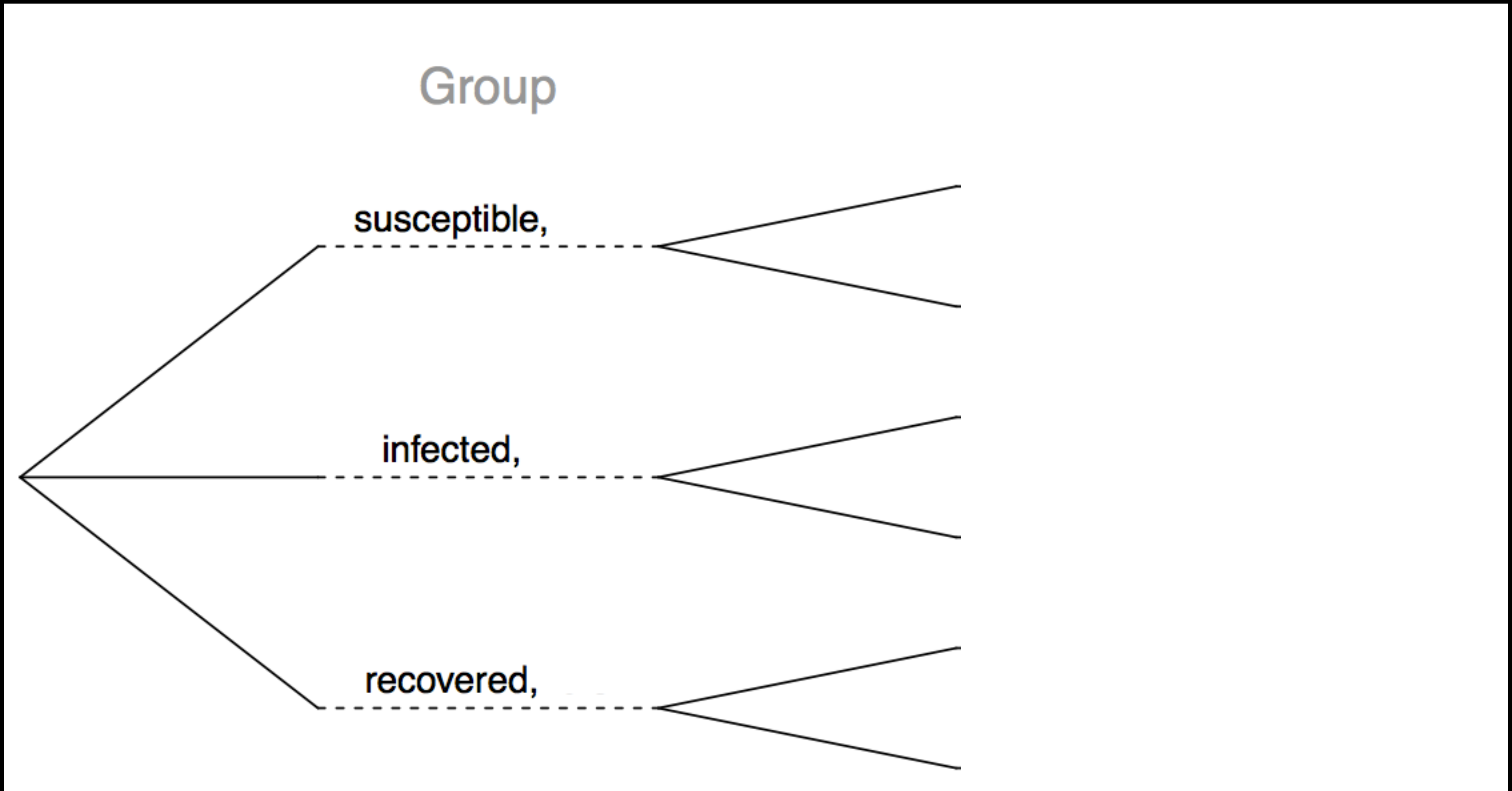
- What we want is: P(infected | +)

Probability of both infected & positive

$$P(inf|+) = \frac{P(inf\ and\ +)}{P(+)}$$

All possibilities for positive tests

Application activity: inverting probabilities (cont.)



$$P(inf|+) =$$

Random variables

As we have been discussing, a random variable is a numeric quantity whose value depends on the outcome of a random event

There are two types of random variables:

Discrete random variables

- Example: Number of credit hours, Difference in number of credit hours this term vs last

Continuous random variables

- Example: Cost of books this term, Difference in cost of books this term vs last

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

Expected value of a discrete random variable

If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) = \mu_x &= x_1 \times p_1 + x_2 \times p_2 + \cdots + x_n \times p_n \\ &= \sum_{i=1}^{n} (x_i \times p_i) \end{aligned} \tag{3.94}$$

Value of thing i

How likely thing i is to happen

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.

Outcome **Probability** **(Outcome X Probability)**

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$

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.

Probability

Outcome **(Outcome X Probability)**

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$

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.

Probability

Outcome **(Outcome X Probability)**

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	$\frac{1}{52}$	$\frac{10}{52}$

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.

Probability

Outcome **(Outcome X Probability)**

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

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.

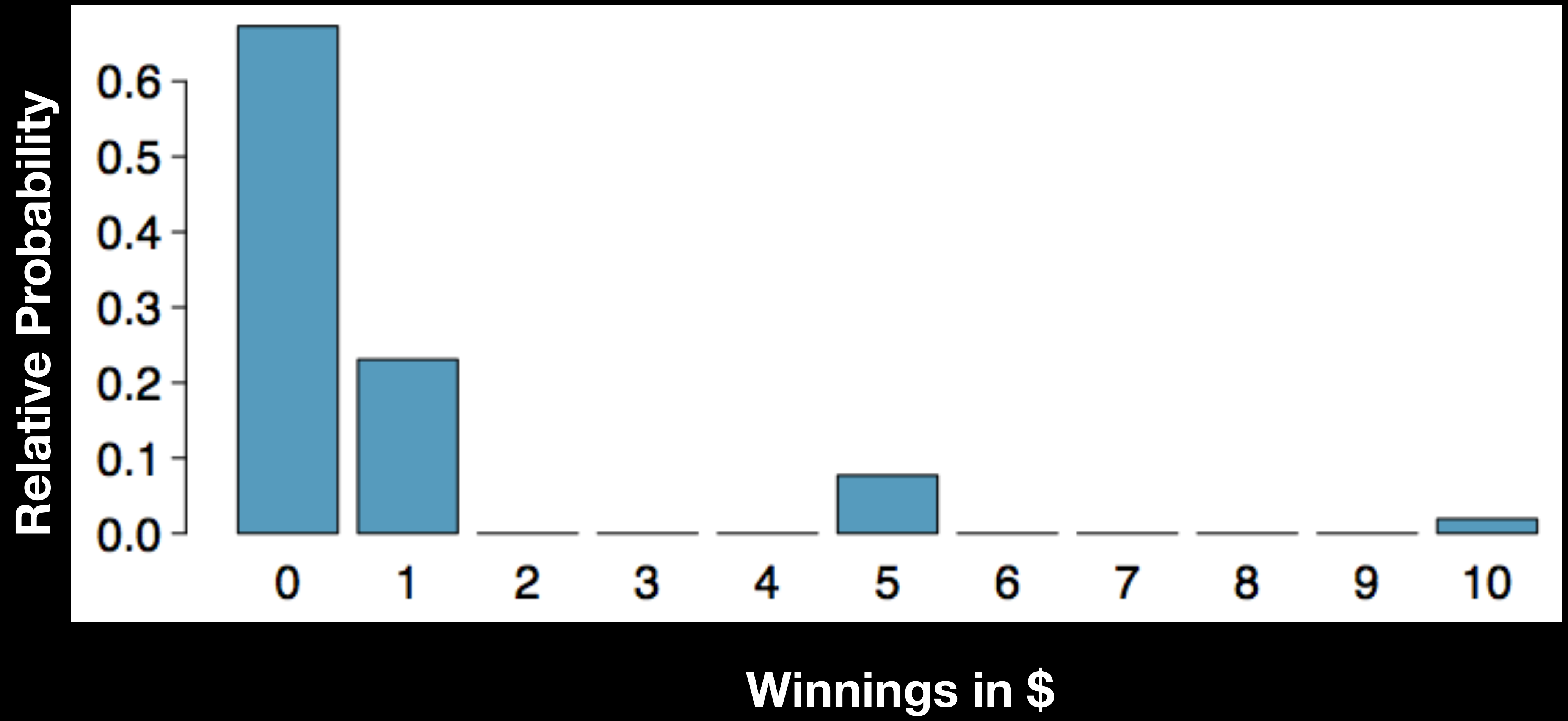
	Outcome	Probability	(Outcome X Probability)
Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

This is about how much money you can expect to make

E(X) = sum(X•P(X))

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.

Variance and standard deviation of a discrete random variable

If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n and expected value $\mu_x = E(X)$, then to find the standard deviation of X, we first find the variance and then take its square root.

$$\begin{aligned} Var(X) = \sigma_x^2 &= (x_1 - \mu_x)^2 \times p_1 + (x_2 - \mu_x)^2 \times p_2 + \cdots + (x_n - \mu_x)^2 \times p_n \\ &= \sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i \end{aligned}$$

$$SD(X) = \sigma_x = \sqrt{\sum_{i=1}^{n} \underbrace{(x_i - \mu_x)}_{\text{Variance or SD}^2\text{ of thing i}}^2 \times p_i} \qquad (3.95)$$

Probability of thing i

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?

Outcome	**Probability**	**(Outcome X Probability)**	**Var**	**Probability X Var**
X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2$	[illegible]

This is what we calculated just before

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.

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?

Outcome	Probability	(Outcome X Probability)	Var	Probability X Var
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	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$

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.

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?

Outcome	Probability	(Outcome X Probability)	Var	Probability X Var
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	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246} = 1.85$

$$SD(X) = \sigma_x = \sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i}$$

The amount we might win from any game can vary by almost \$2 per game, on average!

Practice

A casino game costs \$5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits (or losses) from playing this game? Remember: profit (or loss) = winnings - cost.

(a) a loss of 10¢
(b) a loss of 25¢
(c) a loss of 30¢
(d) a profit of 5¢

Event	Win	Profit: X	$P(X)$	$X \times P(X)$
Red, A♥	500			
Other	0			
				$E(X) =$

Fair game

A fair game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

If those games cost less than their expected payouts, it would mean that the casinos would be losing money on average, and hence they wouldn't be able to pay for all this:

http://www.flickr.com/photos/aigle_dore/5951714693

Expected Value: Real world example

https://projects.fivethirtyeight.com/mortality-rates-united-states/

$$\text{Death rate for cause i} = \frac{\text{\# of people dying from cause in a county}}{\text{\# of people in a county}}$$

Probability you'll die of cause i in a particular county

Could find total country's death rate of particular cause from E(particular cause)

Could find particular county's death rate of all causes E(all death)

More generally: Linear transformations

A linear transformation of a random variables X is given by

$aX + b$

where a and b are some fixed numbers.

The average and SD of a linear transformation can be found as follows:

$$E(aX + b) = a \times E(X) + b$$
$$SD(aX + b) = |\, a \,| \times SD(X)$$

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 of a linear combination of random variables is given by

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

If X and Y are *independent*, then the SD of the linear combination is given by

$$\mathbf{SD(aX + bY) = \sqrt{(a \times SD(X))^2 + (b \times SD(Y))^2}}$$

Example: Calculating the expectation of a linear combination

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

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

Example: Calculating the expectation of a linear combination

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

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

Linear combinations

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

$$\begin{aligned}
\mathbf{SD(5S + 4C)} &= \sqrt{(\mathbf{5} \times \mathbf{SD(S)})^{\mathbf{2}} + (\mathbf{4} \times \mathbf{SD(C)})^{\mathbf{2}}} \\
&= \sqrt{(\mathbf{5} \times \mathbf{1.5})^{\mathbf{2}} + (\mathbf{4} \times \mathbf{2})^{\mathbf{2}}} \\
&= \sqrt{\mathbf{56.25} + \mathbf{64}} \\
&= \mathbf{10.97}
\end{aligned}$$