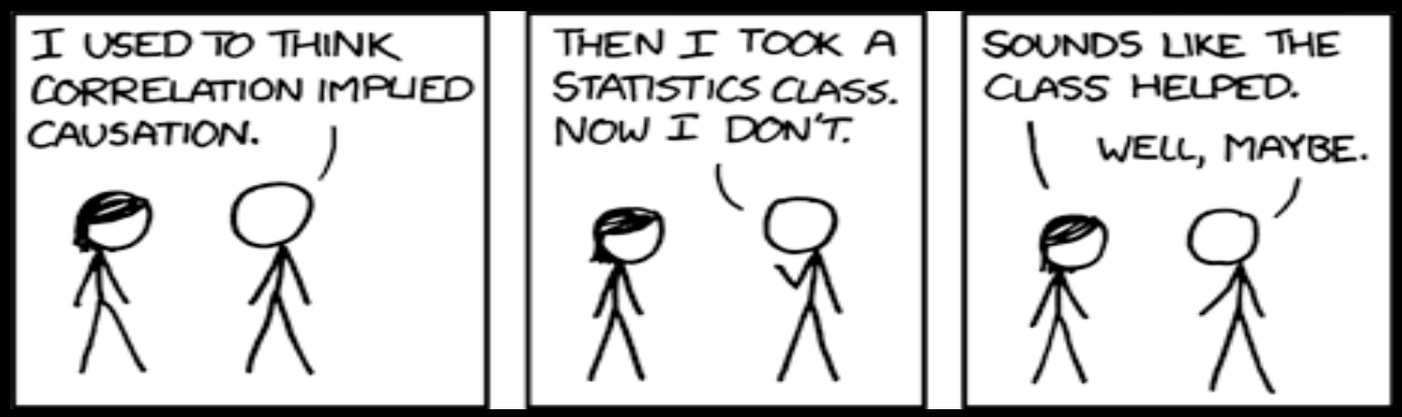
Welcome to Week #4!

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)

Recap

General addition rule

P(A or B) = P(A) + P(B) - P(A and B)

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

Probability distributions

A probability distribution lists all possible events and the probabilities with which they occur.

- The probability distribution for the sex of one kid:

Event	Male	Female
Probability	0.5	0.5

- 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

- The probability distribution for the sexes of two kids:

Event	MM	FF	MF	FM
Probability	0.25	0.25	0.25	0.25

Practice

Sample space is the collection of all possible outcomes of a trial.

- A couple has one kid, what is the sample space for the sex of this kid? S = {M, F}
- A couple has two kids, what is the sample space for the sex of these kids?

Complementary events are two mutually exclusive events whose probabilities that add up to 1.

- A couple has one kid. If we know that the kid is not a boy, what is sex of this kid? { ~~M~~, *F* } Boy and girl are complementary outcomes.
- A couple has two kids, if we know that they are not both girls, what are the possible sex combinations for these kids?

S = { MM, ~~FF~~, FM, MF }

Independence

Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

- Knowing that the coin landed on a head on the first toss does not provide any useful information for determining what the coin will land on in the second toss.
  >> Outcomes of two tosses of a coin are independent.

Product rule for independent events

P(A and B) = P(A) x P(B)
Or more generally, P(A_1 and … and A_k) = P(A_1) x … x P(A_k)

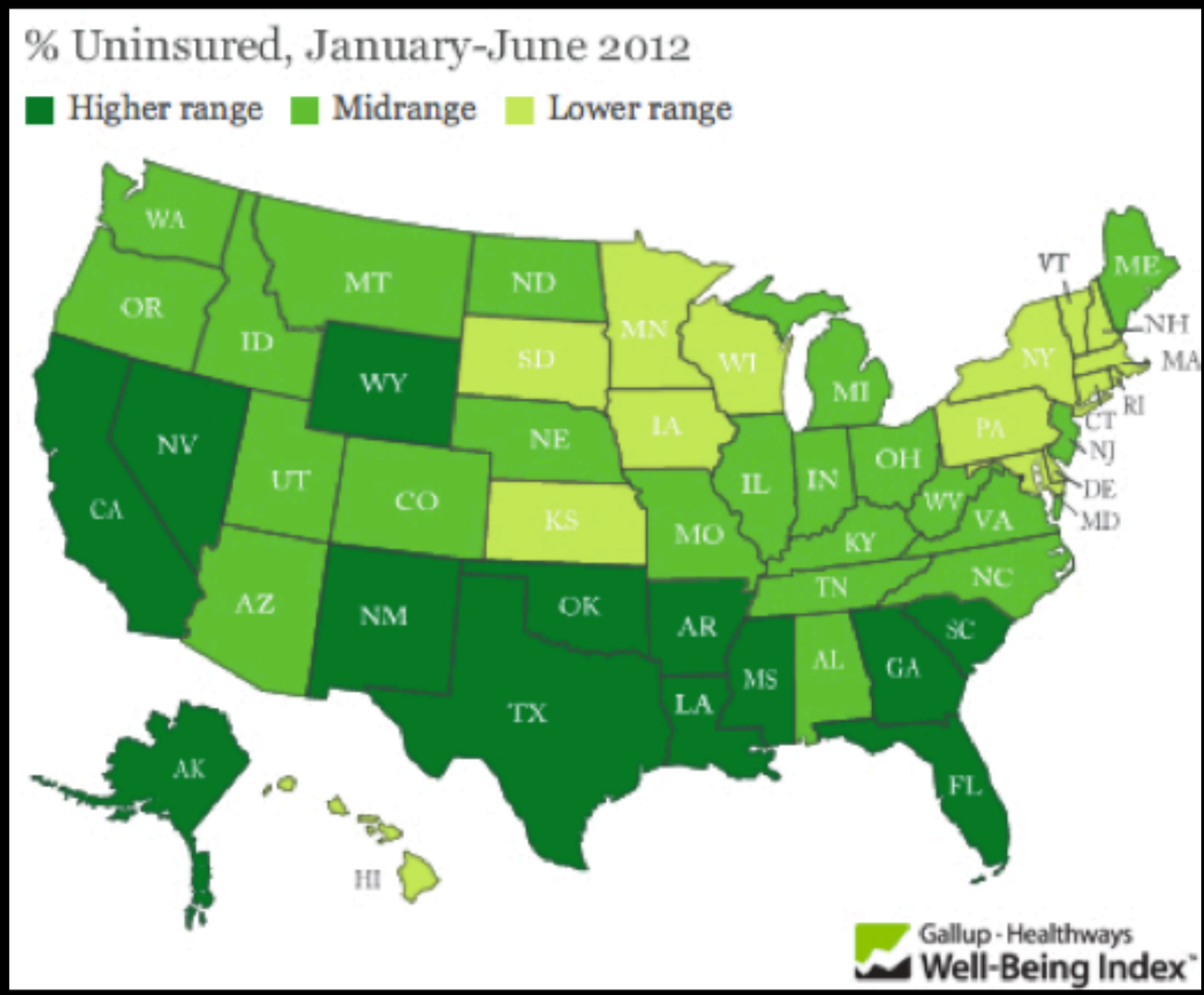
You toss a coin twice, what is the probability of getting two tails in a row?

P(T on the first toss) x P(T on the second toss)
= (1 / 2) x (1 / 2) = 1 / 4

Practice

A recent Gallup poll suggests that 25.5% of Texans do not have health insurance as of June 2012. Assuming that the uninsured rate stayed constant, what is the probability that two randomly selected Texans are both uninsured?

(1) 25.5^2
(2) 0.255^2
(3) 0.255×2
(4) $(1 - 0.255)^2$



http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

Disjoint vs. complementary

Do the sum of probabilities of two disjoint events always add up to 1?

Not necessarily, there may be more than 2 events in the sample space, e.g. party affiliation.

Do the sum of probabilities of two complementary events always add up to 1?

Yes, that's the definition of complementary, e.g. heads and tails.

Practice

Roughly 20% of undergraduates at a university are vegetarian or vegan. What is the probability that, among a random sample of 3 undergraduates, at least one is vegetarian or vegan?

(1) $1 - 0.2 \times 3$

(2) $1 - 0.2^3$

(3) 0.8^3

(4) $1 - 0.8 \times 3$

(5) $1 - 0.8^3$

Relapse

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

Marginal probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed) = 48 / 72 ~ 0.67

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed and desipramine) = 10 / 72 ~ 0.14

Conditional probability

The conditional probability of the outcome of interest A given condition B is calculated as

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

"Probability of event A, given (|) event B"

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(relapse|desipramine) = \frac{P(relapse \text{ and } desipramine)}{P(desipramine)}$$

$$= \frac{10/72}{24/72}$$

$$= \frac{10}{24}$$

$$= 0.42$$

Conditional probability (cont.)

If we know that a patient received the antidepressant (desipramine), what is the probability that they relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapse | desipramine) = 10 / 24 ~ 0.42

P(relapse | lithium) = 18 / 24 ~ 0.75

P(relapse | placebo) = 20 / 24 ~ 0.83

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.

Bayes' Theorem

The conditional probability formula we have seen so far is a special case of the Bayes' Theorem, which is applicable even when events have more than just two outcomes.

Bayes' Theorem

$$P(\textit{outcome A of variable 1} \mid \textit{outcome B of variable 2}) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

where $A_2, \cdots, A_k$ represent all other possible outcomes of variable 1.

See: ThinkBayes2

NOTE: likely to come across this in other classes

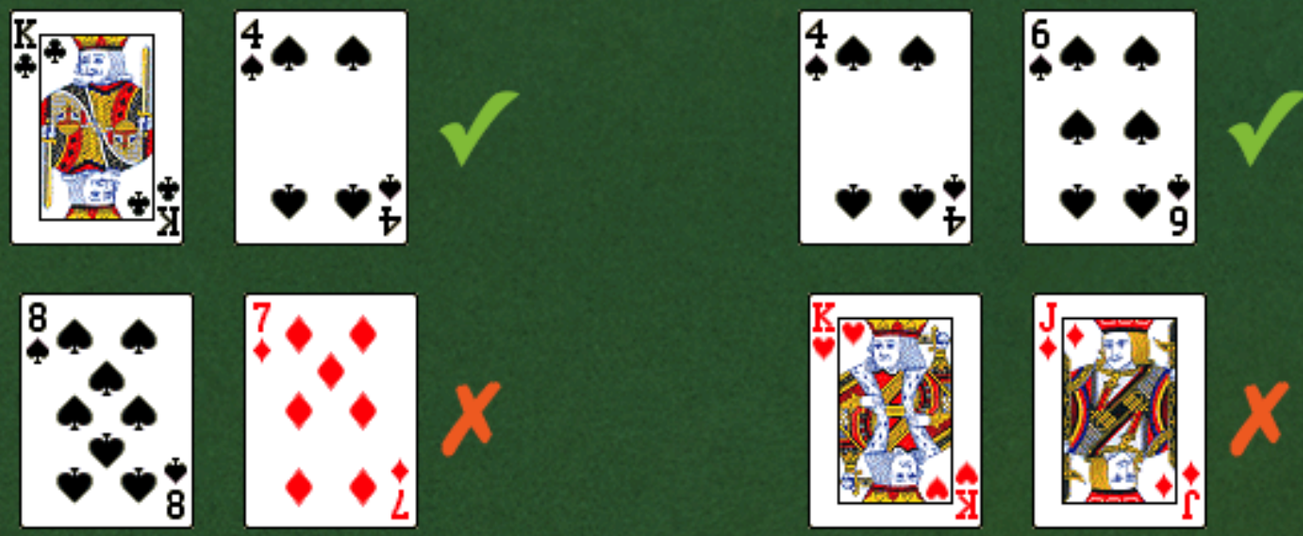
Bank: $171
Level 3

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

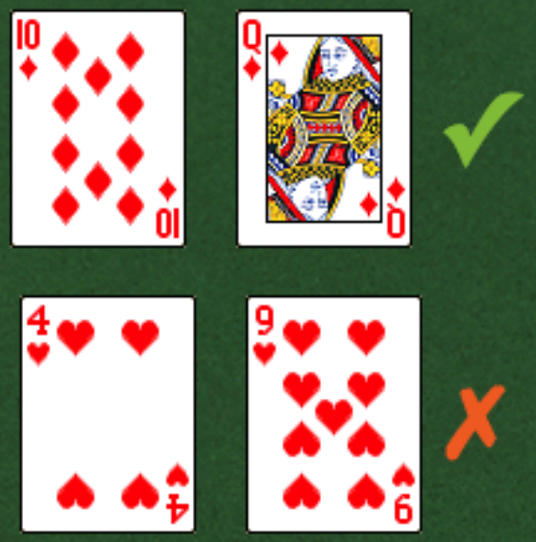
Bank: $260
Level 4

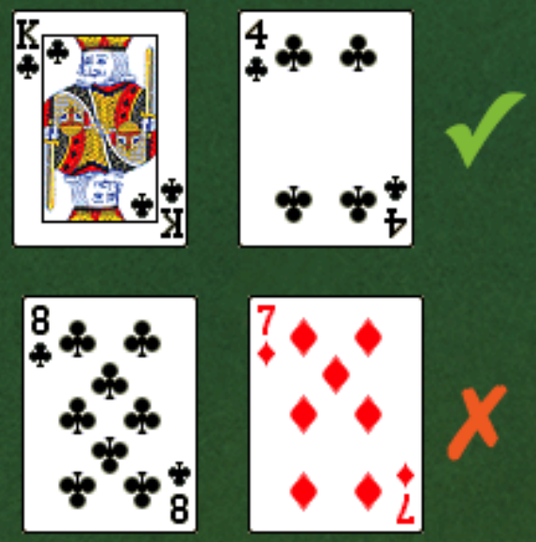
Beat the Odds

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

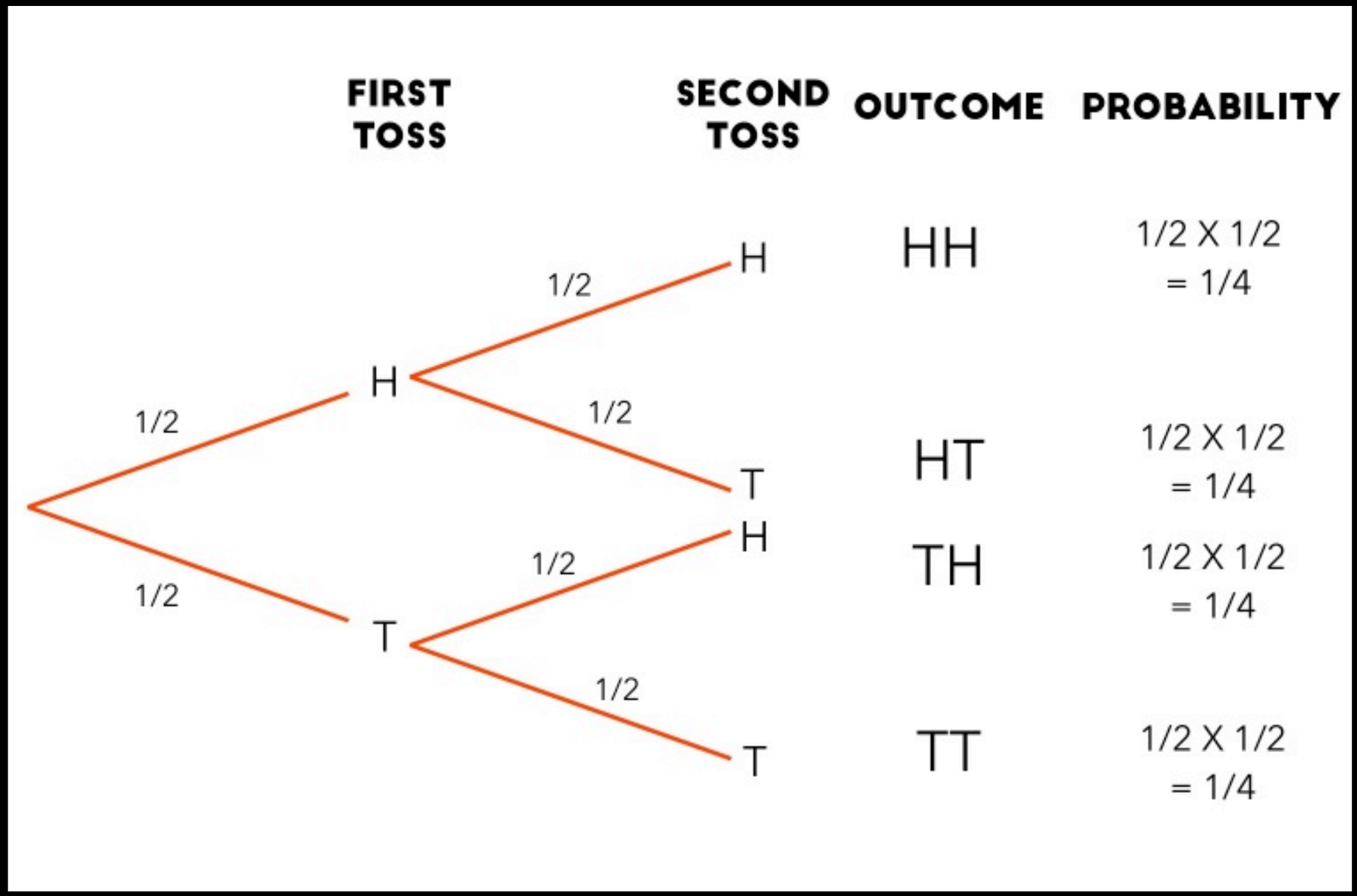
« Back to menu

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

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

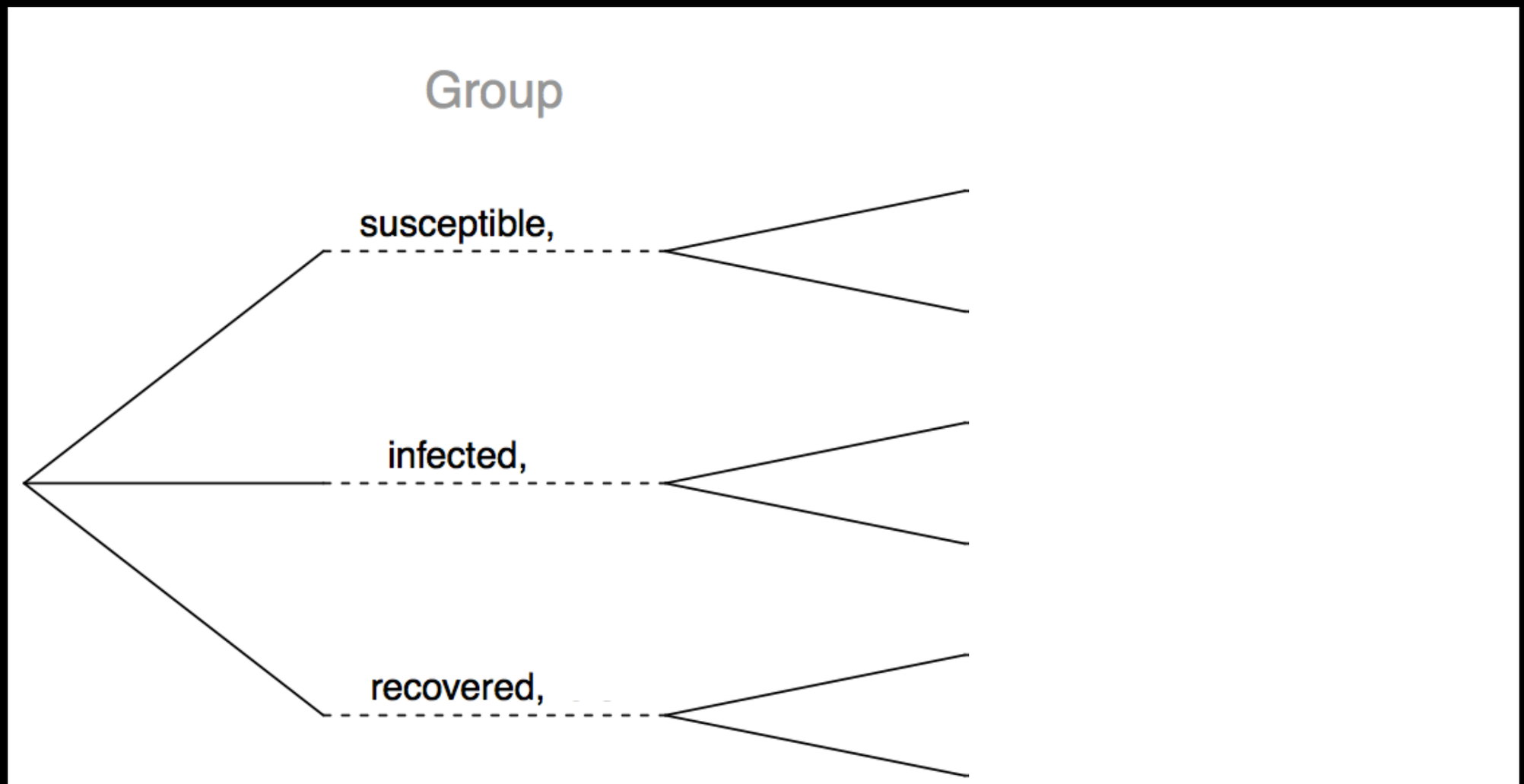
- What we want is: P(infected | +)

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:

$$E(X) = \mu_x = x_1 \times p_1 + x_2 \times p_2 + \cdots + x_n \times p_n$$

$$= \sum_{i=1}^{n} (x_i \times p_i) \qquad (3.94)$$

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.

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

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

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

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

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)**

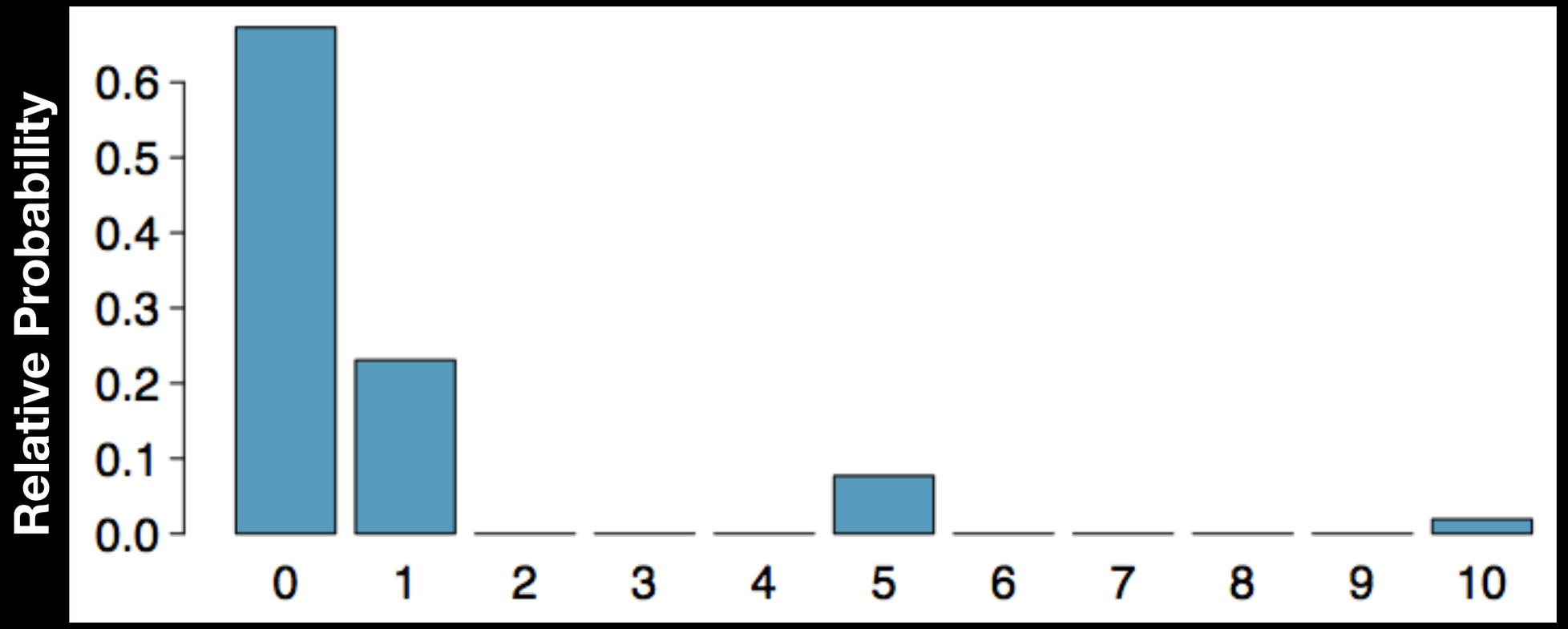
Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

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.

$$Var(X) = \sigma_x^2 = (x_1 - \mu_x)^2 \times p_1 + (x_2 - \mu_x)^2 \times p_2 + \cdots + (x_n - \mu_x)^2 \times p_n$$

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

$$SD(X) = \sigma_x = \sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i} \qquad (3.95)$$

Variance or SD² of thing i

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	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2$	

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 x E(X) + b x 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?

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

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?

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

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?

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