

# Simpson's paradox

## Edward Hugh Simpson

(10 December 1922 – 5 February 2019)

was a British codebreaker, statistician and civil servant.

"The Interpretation of Interaction in Contingency Tables", Journal of the Royal Statistical Society, 1951



Is it possible for one doctor to have a higher success rate than another doctor in every type of treatment he performs but to have a lower overall success rate across all treatment types?



Dr. Hibbert



Dr. Nick



# Simpson's Paradox

	Hibbert heart bandaid	Nick heart bandaid
Success	70	2
Failure	20	8

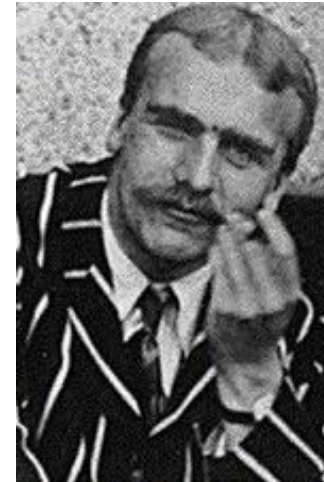
	Hibbert heart bandaid	Nick heart bandaid
Success	10	81
Failure	0	9

Dr. Hibbert: success rate = 80%

Dr. Nick: success rate = 83%

# Simpson's paradox might explain altruism

- Darwinian evolution has a problem with altruism
- “Selfish genes” do not care about others
- J. B. S. Haldane, (1892-1964)  
British geneticist, evolutionary biologist
- When asked if he would give his life to save a drowning brother answered: “No, but I would to save two brothers or eight cousins”
- Altruism in some insect colonies like ants is because they are all genetically similar.



# Altruism in bacteria

- Bacteria live in communities in close proximity to each other
- Individual bugs **spend significant resources** to produce **extracellular molecules**, excrete them outside of the cell to **share with others. That slows their growth**
  - Examples: extracellular enzymes, biofilm components, antimicrobial and anti-immune agents
- **Cheaters have faster growth rate**
  - **They can take over** by not producing any shared molecules
- **Evolutionary paradox: how bacteria can be altruistic?**



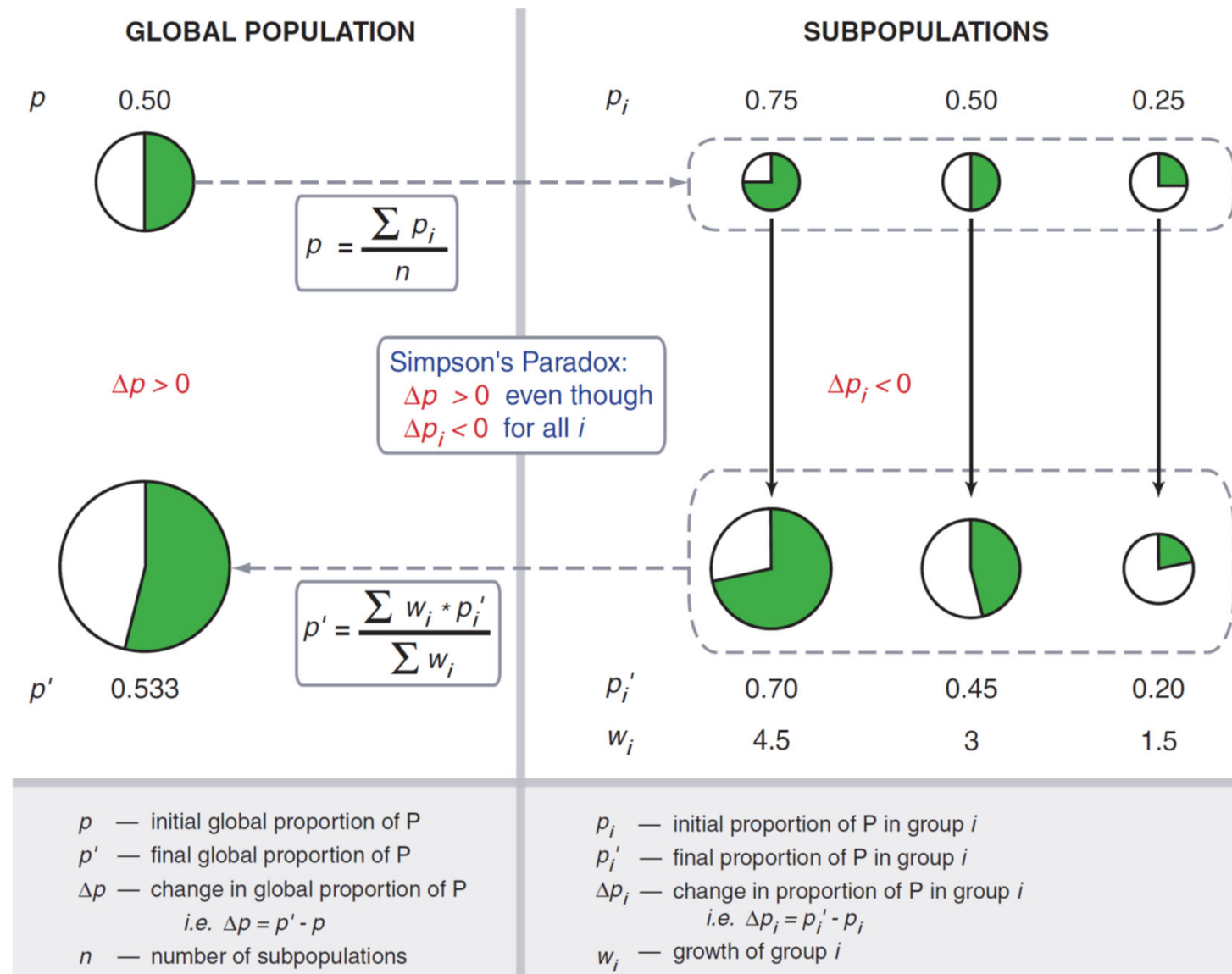


## Simpson's Paradox in a Synthetic Microbial System

John S. Chuang,\* Olivier Rivoire, Stanislas Leibler

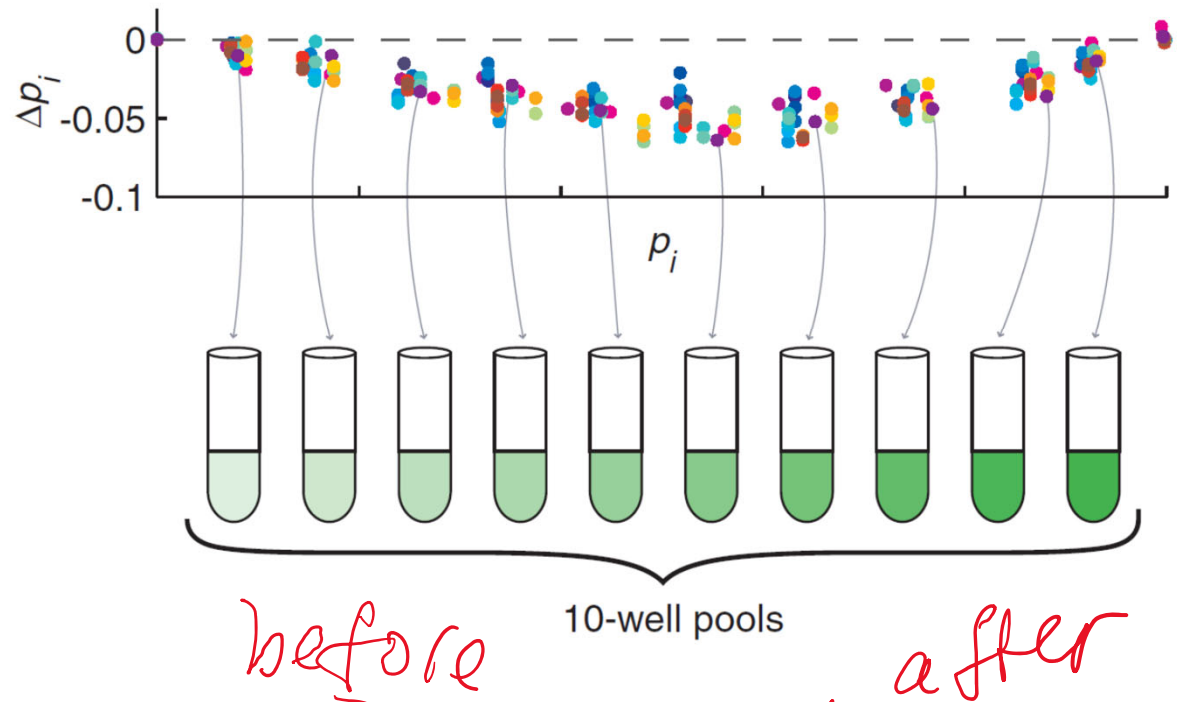
The maintenance of “public” or “common good” producers is a major question in the evolution of cooperation. Because nonproducers benefit from the shared resource without bearing its cost of production, they may proliferate faster than producers. We established a synthetic microbial system consisting of two *Escherichia coli* strains of common-good producers and nonproducers. Depending on the population structure, which was varied by forming groups with different initial compositions, an apparently paradoxical situation could be attained in which nonproducers grew faster within each group, yet producers increased overall. We show that a simple way to generate the variance required for this effect is through stochastic fluctuations via population bottlenecks. The synthetic approach described here thus provides a way to study generic mechanisms of natural selection.

- The common good was a membrane-permeable Rhl autoinducer molecule rewired to activate antibiotic (chloramphenicol; Cm) resistance gene expression.

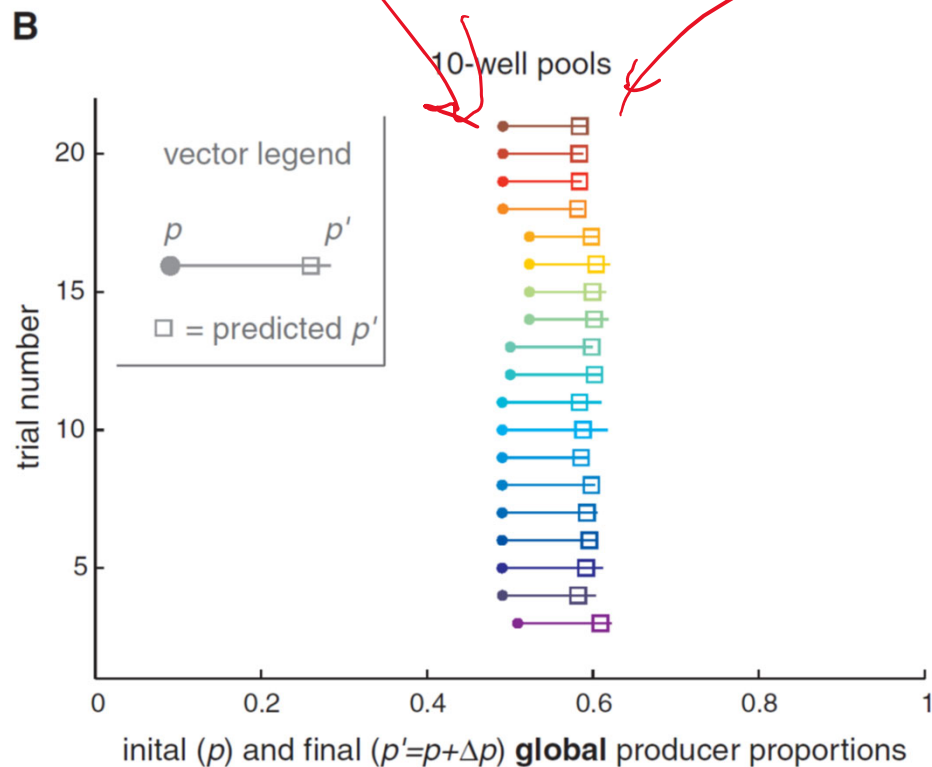




Fraction of altruists in  
each of individual  
test tubes dropped



Yet the overall fraction of  
altruists in  
all test tubes combined  
increased



Credit: XKCD  
comics

# WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS  
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED  
WHY IS THERE A LINE THROUGH HTTPS  
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK  
WHY IS HTTPS IMPORTANT

# QUESTIONS FOUND IN GOOGLE AUTOCOMPLETE



WHY ARE THERE WEEKS  
WHY DO I FEEL DIZZY

WHY DO WHALES JUMP  
WHY ARE WITCHES GREEN  
WHY ARE THERE MIRRORS ABOVE BEDS  
WHY DO I SAY UH  
WHY IS SEA SALT BETTER  
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS  
WHY IS THERE NOT A POKEMON MMO  
WHY IS THERE LAUGHING IN TV SHOWS  
WHY ARE THERE DOORS ON THE FREEWAY  
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING  
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA  
WHY ARE THERE SCARY SOUNDS IN MINECRAFT  
WHY IS THERE KICKING IN MY STOMACH  
WHY ARE THERE TWO SLASHES AFTER HTTP  
WHY ARE THERE CELEBRITIES  
WHY DO SNAKES EXIST  
WHY DO OYSTERS HAVE PEARLS  
WHY ARE DUCKS CALLED DUCKS  
WHY DO THEY CALL IT THE CLAP  
WHY ARE KYLE AND CARTMAN FRIENDS  
WHY IS THERE AN ARROW ON AANG'S HEAD  
WHY ARE TEXT MESSAGES BLUE  
WHY ARE THERE MUSTACHES ON CLOTHES  
WHY ARE THERE MUSTACHES ON CARS  
WHY ARE THERE MUSTACHES EVERYWHERE  
WHY ARE THERE SO MANY BIRDS IN OHIO  
WHY IS THERE SO MUCH RAIN IN OHIO  
WHY IS OHIO WEATHER SO WEIRD

WHY AREN'T ECONOMISTS RICH  
WHY DO AMERICANS CALL IT SOCCER  
WHY ARE MY EARS RINGING  
WHY ARE THERE SO MANY AVENGERS  
WHY ARE THE AVENGERS FIGHTING THE X MEN  
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SWARMS OF GNATS  
WHY IS THERE PHLEGM  
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN  
WHY IS PSYCHIC WEAK TO BUG  
WHY DO CHILDREN GET CANCER  
WHY IS POSEIDON ANGRY WITH ODYSSEUS  
WHY IS THERE ICE IN SPACE

# WHY ARE THERE ANTS IN MY LAPTOP

WHY ARE THERE BRIDESMAIDS  
WHY DO DYING PEOPLE REACH UP  
WHY AREN'T THERE VARICOSE ARTERIES  
WHY ARE OLD KUNGONS DIFFERENT



WHY IS EARTH TILTED  
WHY IS SPACE BLACK  
WHY IS OUTER SPACE SO COLD  
WHY ARE THERE PYRAMIDS ON THE MOON  
WHY IS NASA SHUTTING DOWN



WHY IS THERE AN OWL IN MY BACKYARD  
WHY IS THERE AN OWL OUTSIDE MY WINDOW  
WHY IS THERE AN OWL ON THE DOLLAR BILL  
WHY DO OWLS ATTACK PEOPLE  
WHY ARE AK 47s SO EXPENSIVE  
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE  
WHY ARE THERE GODS  
WHY ARE THERE TWO SPOCKS

WHY ARE DOGS AFRAID OF FIREWORKS  
WHY IS THERE NO KING IN ENGLAND

WHY IS PROGRAMMING SO HARD  
WHY IS THERE A 0 OHM RESISTOR  
WHY DO AMERICANS HATE SOCCER  
WHY DO RHYMES SOUND GOOD  
WHY DO TREES DIE  
WHY IS THERE NO SOUND ON CNN  
WHY AREN'T POKEMON REAL  
WHY AREN'T BULLETS SHARP  
WHY DO DREAMS SEEM SO REAL

WHY ARE THERE TINY SPIDERS IN MY HOUSE  
WHY DO SPIDERS COME INSIDE  
WHY ARE THERE HUGE SPIDERS IN MY HOUSE  
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE  
WHY ARE THERE SPIDERS IN MY ROOM  
WHY ARE THERE SO MANY SPIDERS IN MY ROOM  
WHY DO SPIDER BITES ITCH  
WHY IS DYING SO SCARY

WHY IS THERE NO GPS IN LAPTOPS  
WHY DO KNEES CLICK  
WHY AREN'T THERE E GRADES  
WHY IS ISOLATION BAD  
WHY DO BOYS LIKE ME  
WHY DON'T BOYS LIKE ME  
WHY IS THERE ALWAYS A JAVA UPDATE  
WHY ARE THERE RED DOTS ON MY THIGHS  
WHY IS LYING GOOD



WHY IS MT VESUVIUS THERE  
WHY DO THEY SAY T MINUS  
WHY ARE THERE OBELISKS  
WHY ARE WRESTLERS ALWAYS WET  
WHY ARE OCEANS BECOMING MORE ACIDIC  
WHY IS ARWEN DYING  
WHY AREN'T MY QUAIL LAYING EGGS  
WHY AREN'T MY QUAIL EGGS HATCHING  
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY ARE CIGARETTES LEGAL  
WHY ARE THERE DUCKS IN MY POOL  
WHY IS JESUS WHITE  
WHY IS THERE LIQUID IN MY EAR  
WHY DO Q TIPS FEEL GOOD  
WHY DO GOOD PEOPLE DIE



WHY ARE ULTRASOUNDS IMPORTANT  
WHY ARE ULTRASOUND MACHINES EXPENSIVE  
WHY IS STEALING WRONG



# Monty Hall problem



**Monty Hall** OC, OM (born Monte Halparin)

August 25, 1921 – September 30, 2017

was a Canadian-American game show host, producer, and philanthropist

**Game show “Let’s Make a Deal” aired 1963-now**

# Monty Hall problem

- In *Make a Deal* there are three closed doors. Behind a **random one of these doors** is a car; behind the other two are goats. **The contestant does not know where the car is, but Monty Hall does.**
- After the contestant picks a door Monty **always opens one of the remaining doors**, one he knows does not hide the car. If the contestant has already chosen the door with the car behind, **Monty is equally likely to open either of the two remaining doors.**
- After Monty has shown a goat behind the door that he opens, the **contestant is always given the option to switch doors.**
- What is the probability of winning the car under the switching and non-switching strategies?

Monty Hall problem.  
What strategy  
gives you a better chance  
to win the car?

- A. Better to switch doors
- B. Better not to switch doors
- C. Switching does not matter
- D. The answer depends on the phase of the moon
- E. I don't know

Get your i-clickers

Monty Hall problem.  
What strategy  
gives you a better chance  
to win the car?

A. Better to switch doors

B. Better not to switch doors

C. Switching does not matter

D. The answer depends on the phase of the moon

E. I don't know

Get your i-clickers



# Don't feel bad if you guessed wrong

- When first presented with the Monty Hall problem an overwhelming majority of people assume that each door has an equal probability and conclude that switching does not matter
- Out of 228 subjects in one study, only 13% chose to switch
- Paul Erdős, one of the most prolific mathematicians in history, remained unconvinced until he was shown a computer simulation confirming the predicted result
- Pigeons repeatedly exposed to the problem show that they rapidly learn always to switch, unlike humans

# Solution #1 (intuitive)

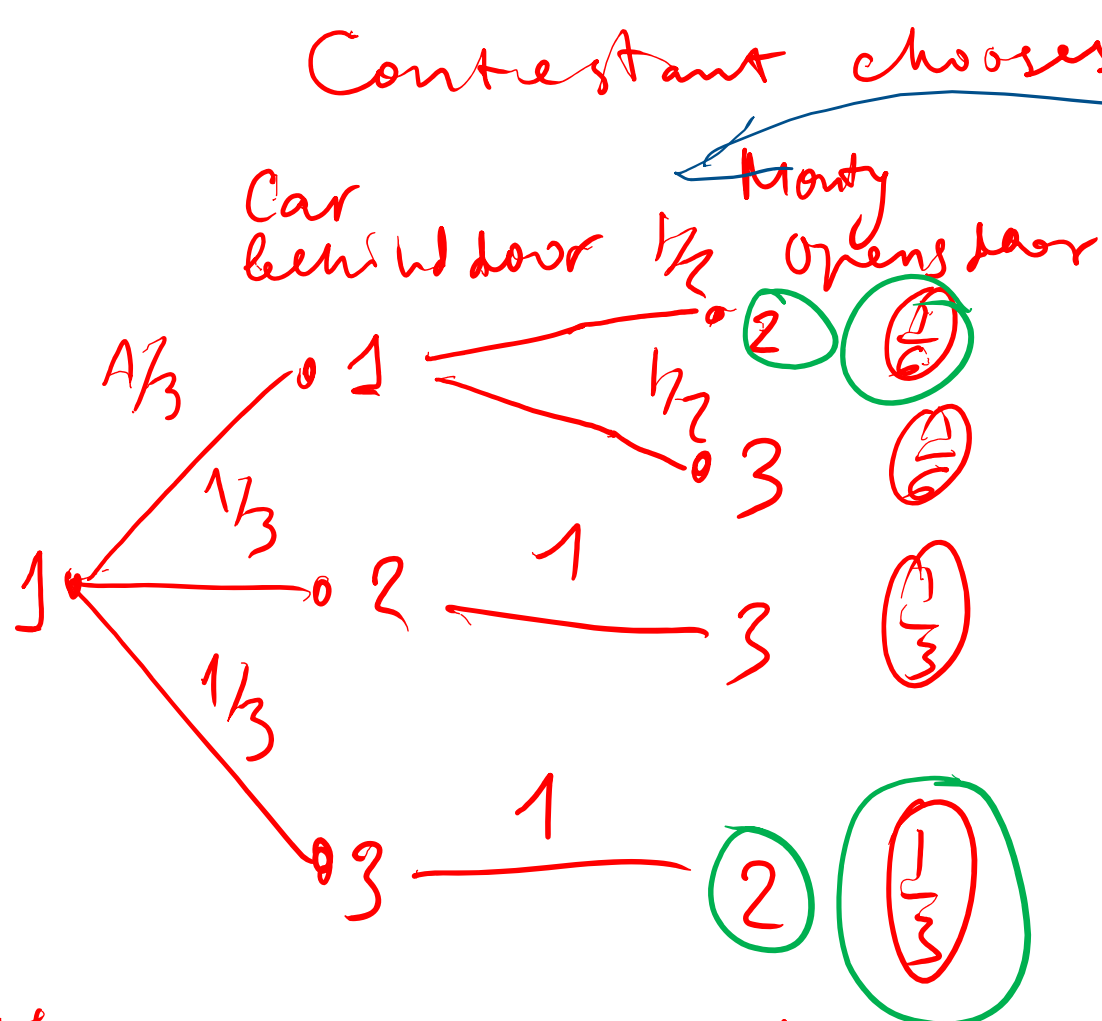
- With **Prob=1/3** you guess the car door right:  
**you loose the car if you switch**
- With **Prob=2/3** you got it wrong and picked a goat door. Then Monty opens another goat door. What is left is the car door.  
**You win the car if you switch!**

## Solution #2.

Tree & conditional probabilities

# Solution #2.

## Tree & conditional probabilities



conditional probability

$$P(\text{Monty opens door 2} \mid \text{car behind door 1}) = \frac{1}{2}$$

If Monty opened door 2

$$P(\text{win by switching}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

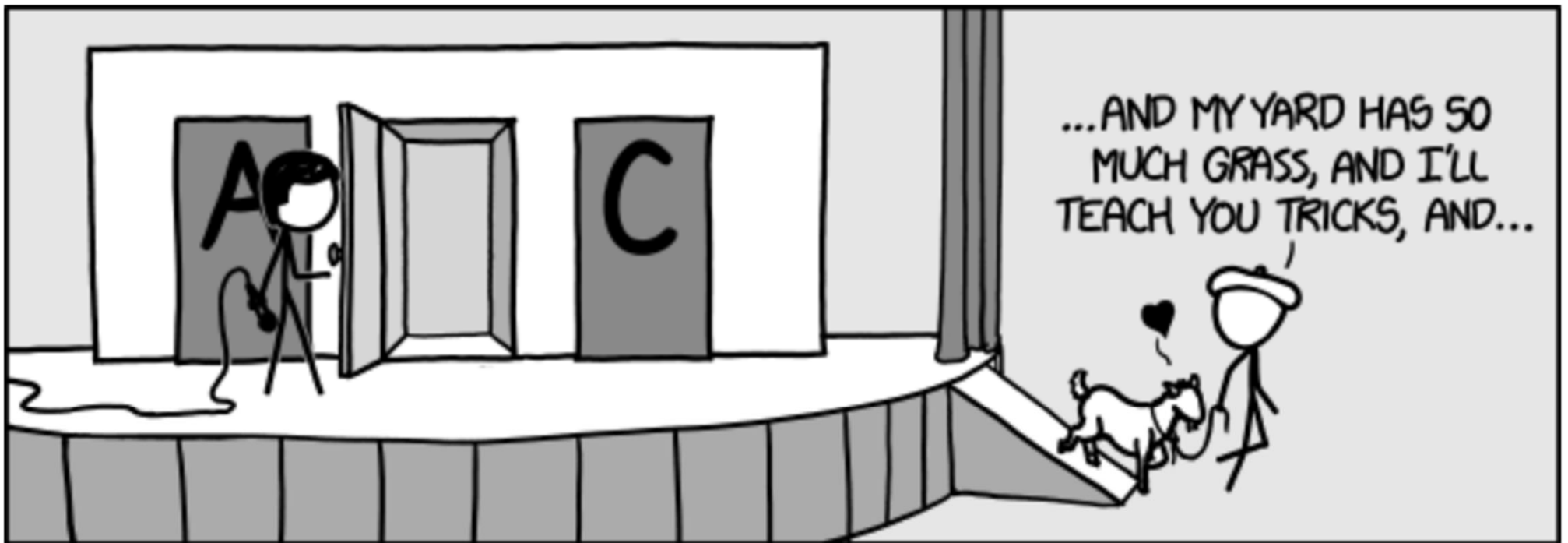
$$P(\text{win by staying}) = \frac{1/6}{1/3 + 1/6} = \frac{1}{3}$$

# A more detailed tree diagram

- Shinyapp website

<https://dacalderon.shinyapps.io/montyhall/>

Thanks to my former BIOE 505 student,  
Alejandra Zeballos Castro, for bringing it to my attention



comic credits: [xkcd](#)

# Let's check the theory by playing the game

Go to

<https://dacalderon.shinyapps.io/montyhall/>

- Tables 1,3,5 will play “switch the door” strategy
- Tables 2,4,6 will play “same door” strategy
- Play at least 30 rounds (more is better)
- In the end we will **add up the numbers from all tables**



		Switch strategy				No switch strategy	
Table		Played	Won	Table		Played	Won
	1	30	18		2	30	8
	3	30	15		4	30	10
	5				6		
		60	33			60	18
P(win)		0.55				0.3	

# Let's check with more random experiments

- `Stats=??;`
- `%set Stats large...`
- `switch_count=0; noswitch_count=0; %set 0 at the beginning`
- `for n = 1:Stats`
- `a = randperm(3); %Monty places two goats and the car at random`
- `%a(1) -goat, a(2) -goat, a(3) - car`
- `i= floor(3.*rand)+1; %you select the door!`
- `% SWITCH STRATEGY`
- `if(i == a(1)) switch_count=switch_count+??; %a(2)-opened, switch to a(3), car!`
- `elseif (i == a(2)) switch_count = switch_count + ??;%a(1) opened, switch to a(3), car!`
- `else switch_count = switch_count + ??; %a(1)/a(2) opened, switch to a(2)/a(1), no car :-(`
- `end`
- `% NO SWITCH STRATEGY`
- `if(i == a(1)) noswitch_count = noswitch_count + ??; %a(2)-opened, no car :-(`
- `elseif (i==a(2)) noswitch_count = noswitch_count + ?? %a(1)-opened, no car :-(`
- `else noswitch_count = noswitch_count + ??; %a(1) or a(2)-opened, car!`
- `endend;`
- `disp('probability to win a car if switched doors=');`
- `disp(num2str(switch_count./??)); %# of cars with switching`
- `disp('probability to win a car if did not switch doors=');`
- `disp(num2str(noswitch_count./??)); %# of cars w/o switching`

# Matlab program

- `B=10000; %set B large...`
- `cars=0; carn=0; %set 0 at the beginning`
- `for i = 1:B`
- `a = randperm(3); %Monty places two goats and the car at random`
- `%a(1) -goat, a(2) -goat, a(3) - car`
- `i= floor(3.*rand)+1; %you select the door!`
- `% SWITCH STRATEGY`
- `if(i == a(1)) cars=cars+1; %a(2)-opened, switch to a(3), car!`
- `elseif (i == a(2)) cars = cars + 1 ;%a(1) opened, switch to a(3), car!`
- `else cars = cars + 0; %a(1)/a(2) opened, switch to a(2)/a(1), no car!`
- `end`
- `% SWITCH STRATEGY`
- `if(i == a(1)) carn = carn + 0; %a(2)-opened, no car`
- `elseif (i==a(2)) carn = carn + 0; %a(1)-opened, no car`
- `else carn = carn + 1; %a(1) or a(2)-opened, car!`
- `end`
- `end;`
- `disp('probability to win a car if switched doors=');`
- `disp(num2str(cars./B)); %# of cars with switching`
- `disp('probability to win a car if did not switch doors=');`
- `disp(num2str(carn./B)); %# of cars w/o switching`

Credit: XKCD  
comics

# WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS  
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED  
WHY IS THERE A LINE THROUGH HTTPS  
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK  
WHY IS HTTPS IMPORTANT

# QUESTIONS FOUND IN GOOGLE AUTOCOMLETE



WHY ARE THERE WEEKS  
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH  
WHY DO AMERICANS CALL IT SOCCER  
WHY ARE MY EARS RINGING  
WHY ARE THERE SO MANY AVENGERS  
WHY ARE THE AVENGERS FIGHTING THE X MEN  
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SWARMS OF GNATS  
WHY IS THERE PHLEGM  
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN  
WHY IS PSYCHIC WEAK TO BUG  
WHY DO CHILDREN GET CANCER  
WHY IS POSEIDON ANGRY WITH ODYSSEUS  
WHY IS THERE ICE IN SPACE

# WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED  
WHY IS SPACE BLACK  
WHY IS OUTER SPACE SO COLD  
WHY ARE THERE PYRAMIDS ON THE MOON  
WHY IS NASA SHUTTING DOWN



WHY IS THERE AN OWL IN MY BACKYARD  
WHY IS THERE AN OWL OUTSIDE MY WINDOW  
WHY IS THERE AN OWL ON THE DOLLAR BILL  
WHY DO OWLS ATTACK PEOPLE  
WHY ARE AK 47s SO EXPENSIVE  
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE  
WHY ARE THERE GODS  
WHY ARE THERE TWO SPOCKS

WHY ARE DOGS AFRAID OF FIREWORKS  
WHY IS THERE NO KING IN ENGLAND

WHY DO WHALES JUMP  
WHY ARE WITCHES GREEN  
WHY ARE THERE MIRRORS ABOVE BEDS  
WHY DO I SAY UH  
WHY IS SEA SALT BETTER  
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS  
WHY IS THERE NOT A POKEMON MMO  
WHY IS THERE LAUGHING IN TV SHOWS  
WHY ARE THERE DOORS ON THE FREEWAY  
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING  
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA  
WHY ARE THERE SCARY SOUNDS IN MINECRAFT  
WHY IS THERE KICKING IN MY STOMACH  
WHY ARE THERE TWO SLASHES AFTER HTTP  
WHY ARE THERE CELEBRITIES  
WHY DO SNAKES EXIST  
WHY DO OYSTERS HAVE PEARLS  
WHY ARE DUCKS CALLED DUCKS  
WHY DO THEY CALL IT THE CLAP  
WHY ARE KYLE AND CARTMAN FRIENDS  
WHY IS THERE AN ARROW ON AANG'S HEAD  
WHY ARE TEXT MESSAGES BLUE  
WHY ARE THERE MUSTACHES ON CLOTHES  
WHY ARE THERE MUSTACHES ON CARS  
WHY ARE THERE MUSTACHES EVERYWHERE  
WHY ARE THERE SO MANY BIRDS IN OHIO  
WHY IS THERE SO MUCH RAIN IN OHIO  
WHY IS OHIO WEATHER SO WEIRD

WHY ARE THERE MALE AND FEMALE BIKES  
WHY ARE THERE TINY SPIDERS IN MY HOUSE  
WHY DO SPIDERS COME INSIDE  
WHY ARE THERE HUGE SPIDERS IN MY HOUSE  
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE  
WHY ARE THERE SPIDERS IN MY ROOM  
WHY ARE THERE SO MANY SPIDERS IN MY ROOM  
WHY DO SPIDER BITES ITCH  
WHY IS DYING SO SCARY



WHY ARE THERE BRIDESMAIDS  
WHY DO DYING PEOPLE REACH UP  
WHY AREN'T THERE VARICOSE ARTERIES  
WHY ARE OLD KUNGONS DIFFERENT  
WHY IS THERE HELL IF GOD FORGIVES  
WHY IS THERE NO GPS IN LAPTOPS  
WHY DO KNEES CLICK  
WHY AREN'T THERE E GRADES  
WHY IS ISOLATION BAD  
WHY DO BOYS LIKE ME  
WHY DON'T BOYS LIKE ME  
WHY IS THERE ALWAYS A JAVA UPDATE  
WHY ARE THERE RED DOTS ON MY THIGHS  
WHY IS LYING GOOD



WHY IS MT VESUVIUS THERE  
WHY DO THEY SAY T MINUS  
WHY ARE THERE OBELISKS  
WHY ARE WRESTLERS ALWAYS WET  
WHY ARE OCEANS BECOMING MORE ACIDIC  
WHY IS ARWEN DYING  
WHY AREN'T MY QUAIL LAYING EGGS  
WHY AREN'T MY QUAIL EGGS HATCHING  
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA



WHY ARE CIGARETTES LEGAL  
WHY ARE THERE DUCKS IN MY POOL  
WHY IS JESUS WHITE  
WHY IS THERE LIQUID IN MY EAR  
WHY DO Q TIPS FEEL GOOD  
WHY DO GOOD PEOPLE DIE  
WHY ARE ULTRASOUNDS IMPORTANT  
WHY ARE ULTRASOUND MACHINES EXPENSIVE  
WHY IS STEALING WRONG

WHY IS PROGRAMMING SO HARD  
WHY IS THERE A 0 OHM RESISTOR  
WHY DO AMERICANS HATE SOCCER  
WHY DO RHYMES SOUND GOOD  
WHY DO TREES DIE  
WHY IS THERE NO SOUND ON CNN  
WHY AREN'T POKEMON REAL  
WHY AREN'T BULLETS SHARP  
WHY DO DREAMS SEEM SO REAL

# Discrete Probability Distributions

# Random Variables

- A variable that associates a number with the outcome of a **random experiment** is called a **random variable**.
- Notation: **random variable** is denoted by an uppercase letter, such as *X*. After the experiment is conducted, the **measured value** is denoted by a **lowercase letter**, such as *x*. Both *X* and *x* are shown in italics, e.g.,  $P(X=x)$ .

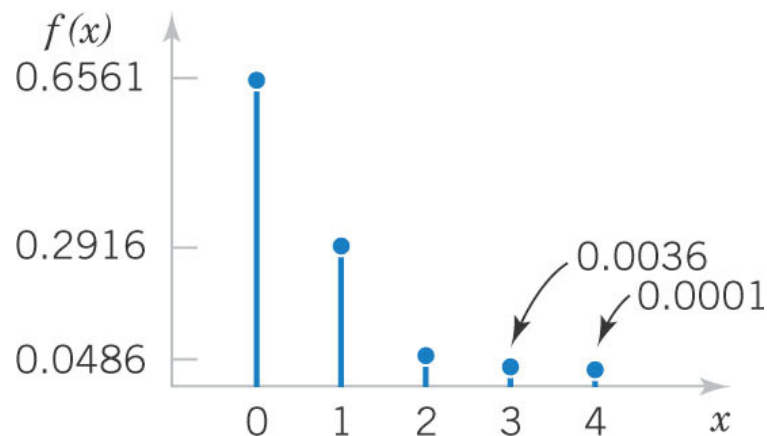


# Continuous & Discrete Random Variables

- A **discrete random variable** is usually integer number
  - $N$  - the number of p53 proteins in a cell
  - $D$  - the number of nucleotides different between two sequences
- A **continuous random variable** is a real number
  - $C=N/V$  – the concentration of p53 protein in a cell of volume  $V$
  - Percentage  $(D/L)*100\%$  of different nucleotides in protein sequences of different lengths  $L$   
(depending on the set of  $L$ 's may be discrete but dense)

# Probability Mass Function (PMF)

- I want to **compare all 4-mers** in a pair of human genomes
- **$X$**  – random variable: the number of nucleotide differences in a given 4-mer
- **Probability Mass Function:**  $f(x)$  or  $P(X=x)$  – the probability that the # of SNPs is **exactly equal to  $x$**



Probability Mass Function for the # of mismatches in 4-mers

$P(X=0) =$	0.6561
$P(X=1) =$	0.2916
$P(X=2) =$	0.0486
$P(X=3) =$	0.0036
$P(X=4) =$	0.0001
$\sum_x P(X=x) =$	1.0000

# Cumulative Distribution Function (CDF)

$x$	$P(X=x)$	$P(X \leq x)$	$P(X > x)$
-1	0.0000	0.0000	1.0000
0	0.6561	0.6561	0.3439
1	0.2916	0.9477	0.0523
2	0.0486	0.9963	0.0037
3	0.0036	0.9999	0.0001
4	0.0001	1.0000	0.0000

Cumulative Distribution Function CDF:  $F(x) = P(X \leq x)$

Example:

$$F(3) = P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0.9999$$

Complementary Cumulative Distribution Function  
(tail distribution) or CCDF:  $F_{>}(x) = P(X > x)$

$$\text{Example: } F_{>}(0) = P(X > 0) = 1 - P(X \leq 0) = 1 - 0.6561 = 0.3439$$

# Mean or Expected Value of $X$

The **mean** or **expected value** of the discrete random variable  $X$ , denoted as  $\mu$  or  $E(X)$ , is

$$\mu = E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x)$$

- **The mean** = the weighted average of all possible values of  $X$ . It represents its “center of mass”
- The **mean may, or may not**, be an **allowed value of  $X$**
- It is also called the **arithmetic mean** (to distinguish from e.g. the **geometric mean** discussed later)
- **Mean may be infinite** if  $X$  any integer and tail  $P(X=x) > c/x^2$



Outcomes of 6 random experiments

0, 1, 0, 0, 2, 1

$$\text{Mean} = \frac{0 + 1 + 0 + 0 + 2 + 1}{6} =$$

$$= \frac{3 \times 0 + 2 \times 1 + 1 \times 2}{6} =$$

$$= 0 \times \frac{3}{6} + 1 \times \frac{2}{6} + 2 \times \frac{1}{6} = \sum_{x=0}^2 x P(X=x)$$





$$\bullet E[X] = \sum_x x \cdot P(X=x)$$

$$\bullet E[X^2] = \sum_x x^2 \cdot P(X=x)$$

$$\bullet E[a \cdot X + b \cdot X^2] = \sum_x (a x + b x^2) \cdot P(X=x) \\ = a \cdot \sum_x x P(X=x) + b \sum_x x^2 P(X=x)$$

$$\bullet E[e^X] = \sum_x e^x P(X=x)$$



Variance  $V(X)$ : Square  
of a typical deviation from  
the mean  $\mu = E(X)$   
 $V(X) = \sigma^2$ , where  $\sigma$  is called  
Standard deviation

$$\begin{aligned}\sigma^2 &= V(X) = E((X - \mu)^2) = \\ &= E(X^2 - 2\mu X + \mu^2) = E(X^2) - \\ &- 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = \\ &= E(X^2) - \mu^2 = E(X^2) - (E(X))^2\end{aligned}$$

# Variance of a Random Variable

If  $X$  is a discrete random variable with probability mass function  $f(x)$ ,

$$E[h(X)] = \sum_x h(x) \cdot P(X = x) = \sum_x h(x) f(x) \quad (3-4)$$

If  $h(x) = (X - \mu)^2$ , then its expectation,  $V(x)$ , is the **variance of  $X$** .

$\sigma = \sqrt{V(x)}$ , is called **standard deviation of  $X$**

$\sigma^2 = V(X) = \sum_x (x - \mu)^2 f(x)$  is the **definitional** formula

$$= \sum_x (x^2 - 2\mu x + \mu^2) f(x)$$

$$= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x)$$

$$= \sum_x x^2 f(x) - 2\mu^2 + \mu^2$$

$$= \sum_x x^2 f(x) - \mu^2 \text{ is the } \mathbf{computational} \text{ formula}$$

**Variance can be infinite**  
if  $X$  can be any integer  
and tail of  $P(X=x) \geq c/x^3$

# Skewness of a random variable

- Want to quantify **how asymmetric** is the **distribution around the mean?**
- Need any **odd moment**:  $E[(X-\mu)^{2n+1}]$
- **Cannot** do it with the **first moment**:  $E[X-\mu]=0$
- Normalized 3-rd moment is **skewness**:  $\gamma_1 = E[(X-\mu)^3]/\sigma^3$
- Skewness **can be infinite** if  $X$  takes unbounded integer values and tail  $P(X=x) \geq c/x^4$

# Geometric mean of a random variable

- Useful for **very broad distributions** (many orders of magnitude)?
- Mean may be dominated by **very unlikely** but **very large events**. Think of a **lottery**
- **Exponent of the mean of  $\log X$ :**  
*Geometric mean =  $\exp(E[\log X])$*
- Geometric mean usually **is not infinite**

# Summary: Parameters of a Probability Distribution

- **Probability Mass Function (PMF):**  $f(x)=\text{Prob}(X=x)$
- **Cumulative Distribution Function (CDF):**  $F(x)=\text{Prob}(X\leq x)$
- **Complementary Cumulative Distribution Function (CCDF):**  
 $F_{>}(x)=\text{Prob}(X>x)$
- The **mean,  $\mu=E[X]$** , is a measure of the **center of mass of a random variable**
- The **variance,  $V(X)=E[(X-\mu)^2]$** , is a measure of the **dispersion** of a random variable **around its mean**
- The **standard deviation,  $\sigma=[V(X)]^{1/2}$** , is another measure of the **dispersion** around mean. Has the same units as  $X$
- The **skewness,  $\gamma_1=E[(X-\mu)^3/\sigma^3]$** , a measure of asymmetry around mean
- The **geometric mean,  $\exp(E[\log X])$**  is useful for very broad distributions