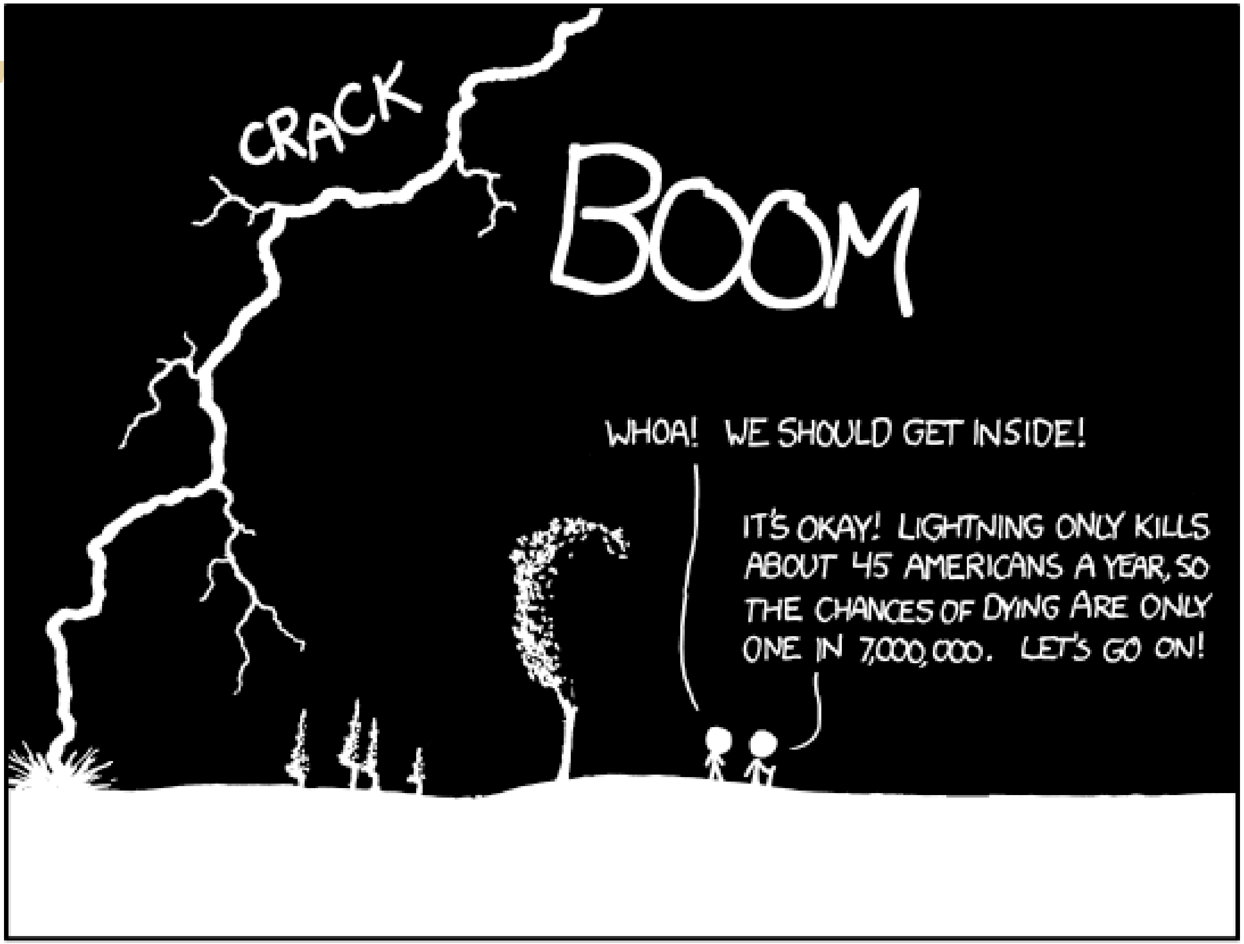
UNIVERSITY of WASHINGTON

Data Science UW
Methods for Data Analysis

Probability and More on Distributions
Lecture 2
Stephen Elston



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Topics

> Review
> Counting
> Axioms of Probability
> Probability Examples
> Conditional Probability
> Simulation

Review

Summary statistics

- Sample mean = μ = sum(x_i)/n
- sample var = σ = sum((μ – x_i)^2) / (n – 1)
- Sample std = sqrt(σ)
- Standard error of the sample mean = se = std/ sqrt(n)

Review

Data exploration and visualization

- Develop understanding of relations in data set
- Use multiple views
- Iterative process
  - Try lots of things
  - Fail lots and fast
  - Find what works

Counting

> Combinatorics of the biggest areas of mathematics.

> Example:
  - Subway has 4 bread choices, 5 meat choices, 4 toppings. How many sandwich combinations?
  - How many different 4-beer tasters can I have in a bar with 10 beers on tap?

> Solve these using the 'Multiplication Principle'.
  - Subway Problem:

$$\underset{\text{(\# of breads)}}{\underline{\quad 4 \quad}} * \underset{\text{(\# of meats)}}{\underline{\quad 5 \quad}} * \underset{\text{(\# of toppings)}}{\underline{\quad 4 \quad}} = 80$$

  - Beer Problem:

$$\underset{\text{(\# for 1}^{\text{st}}\text{ beer)}}{\underline{\quad 10 \quad}} * \underset{\text{(\# for 2}^{\text{nd}}\text{ beer)}}{\underline{\quad 9 \quad}} * \underset{\text{(\# for 3}^{\text{rd}}\text{ beer)}}{\underline{\quad 8 \quad}} * \underset{\text{(\# for 4}^{\text{th}}\text{ beer)}}{\underline{\quad 7 \quad}} = 5{,}040$$

Multiplication Principle

- If there are A ways of doing task a, and B ways of doing task b, then there are A*B ways of completing both tasks.
- Example:
  - If I have 5 books, how many ways can I *order* them on the bookshelf?

5 choices	*	4 choices	*	3 choices	*	2 choices	*	1 choice
$\uparrow$		$\uparrow$		$\uparrow$		$\uparrow$		$\uparrow$
1^{st} book		2^{nd} book		3^{rd} book		4^{th} book		5^{th} book

= 5 factorial = 5! = 120

Factorials

- Factorials
  - Count # ways to order N things = N!
- Factorials get VERY LARGE quickly.
  - 21! Is larger than the biggest long-int in 64 bit.
    - 21! = 5.1E19
    - Biggest long int (64 bit) = 9.2E18
  - Fun fact, every 52 card shuffle is highly likely to be the only time that shuffle has ever occurred.

Counting Subgroups

> Revisit: 10 beers on tap, need a sample of 4 different beers.
> Let's assume order matters, i.e., Amber-Stout-Porter-Red is different from Red-Porter-Stout-Amber.
> Use 'Permutations' (pick):

$$10 * 9 * 8 * 7 = \frac{10!}{6!} = \frac{10!}{(10-4)!} = 10P4 = P(10,4)$$

Counting Subgroups

> Now, Let's assume order doesn't matter.
> Use 'Combinations' (choose):

$$\frac{10*9*8*7=\frac{10!}{6!}=\frac{10!}{(10-4)!}=10P4=P(10,4)}{(\#\ of\ orderings\ of\ 4\ beers)=4!}$$

$$=\frac{10!}{4!\,(10-4)!}=10C4=C(10,4)=\binom{10}{4}$$

More on Combinations

- Combinations appear on the Pascal's Triangle!
- C(N,x) appears on the Nth row, xth number (starting at 0)

$\binom{10}{4}$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1
1 11 55 165 330 462 462 330 165 55 11 1
1 12 66 220 495 792 924 792 495 220 66 12 1

Counting Examples

> There are 10 Light beers on tap, and 10 Dark beers on tap, how many ways can I get a 4-beer sampler that contains exactly 1 light beer? (ordering doesn't matter)

$$\frac{(\#\ of\ ways\ for\ light\ beer) \cdot (\#\ of\ ways\ for\ dark\ beer)}{(\#\ of\ ways\ to\ order\ 1L\ and\ 3D)}$$

$$\frac{(10) \cdot \binom{10}{3}}{4} = \frac{10 * 120}{4} = 300$$

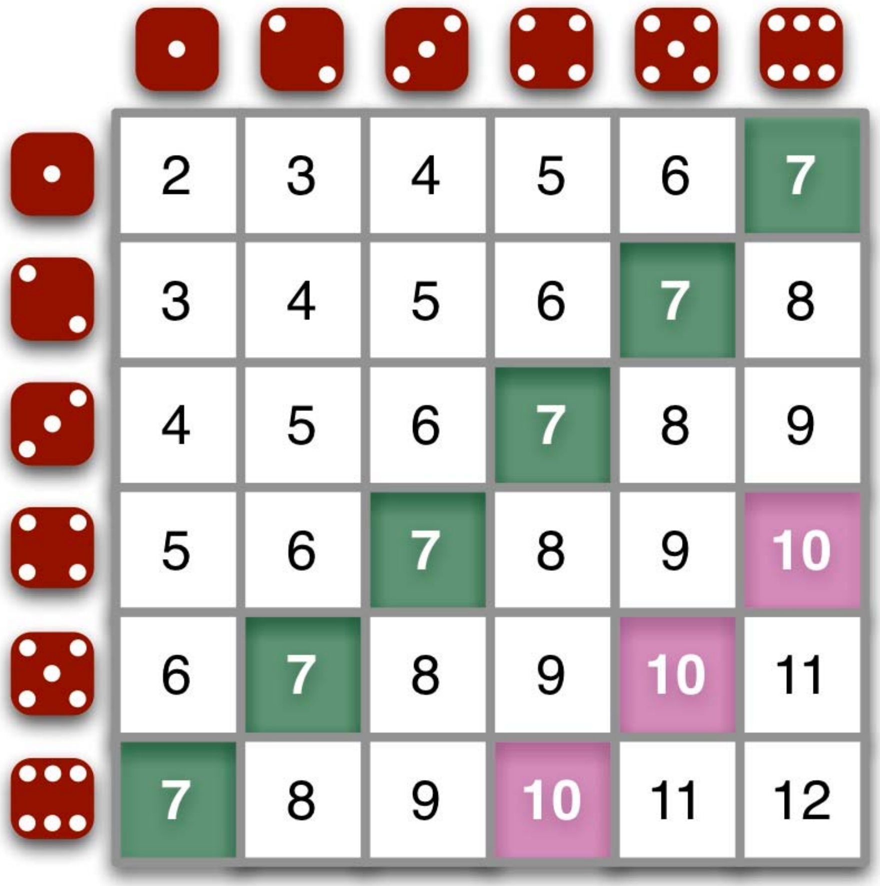
Counting Examples

- 6:5 Blackjack is dealt with a 6 shoe deck (52*6=312 cards). How many ways can someone get dealt two rank 10 cards?

$$\binom{6decks * 4ranks * 4suits}{2} = \binom{96}{2} = \frac{96!}{2!\,(94!)} = \frac{96 * 95}{2} = 4560$$

Counting Examples

> How many ways can two dice be rolled to get a sum of 10?



Counting in R

- expand.grid() – function that creates a data frame from all combinations of vectors supplied.
- R-demo

Probability

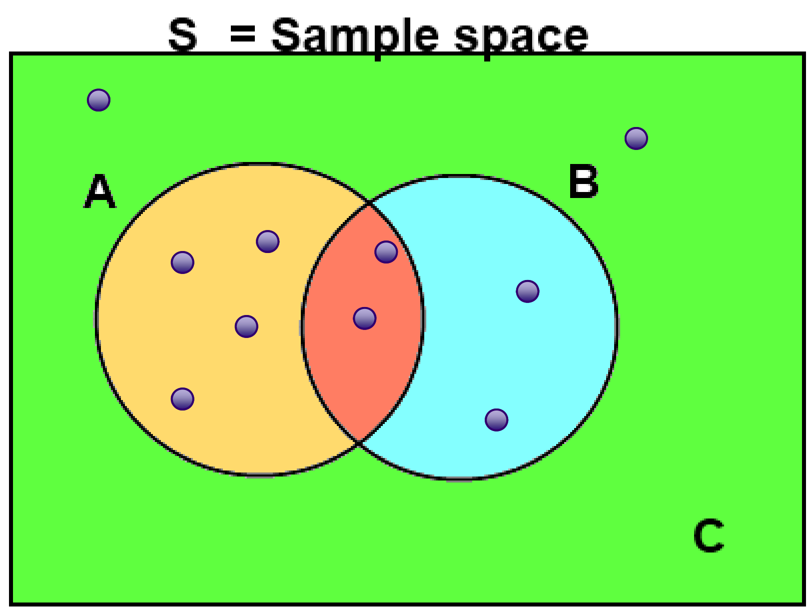
> The Probability of an event, A, is the number of ways A can occur, divided by the number of total possible outcomes in our Sample Space, S.

$$P(A) = \frac{N(A)}{N(S)}$$

> If • is an event, then

$$P(A) = \frac{6}{10} = \frac{3}{5}$$

$$P(B) = \frac{4}{10} = \frac{2}{5}$$

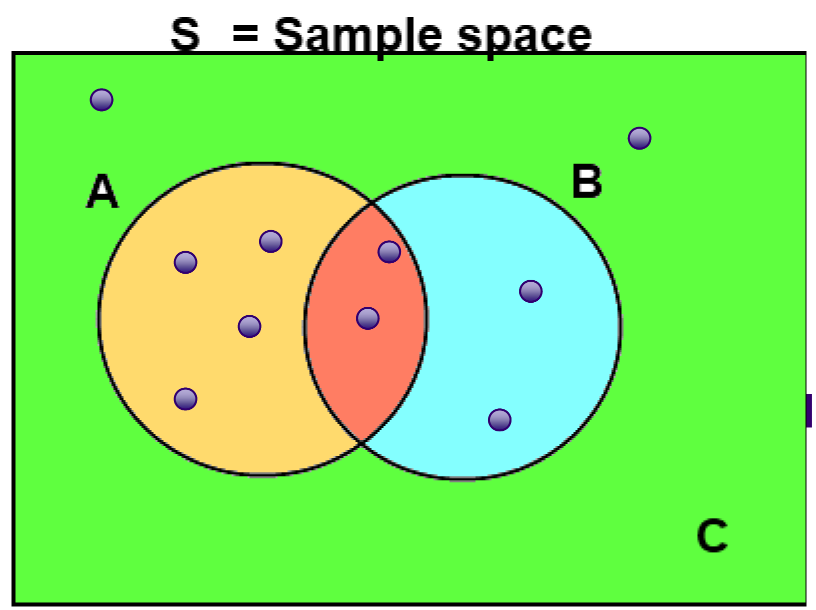


Probability

- If  is an event, then
  - Intersection: $P(A \cap B) = \frac{2}{10} = \frac{1}{5}$
  - Union: $P(A \cup B) = \frac{8}{10} = \frac{4}{5}$
  - Negation: $P(A') = \frac{4}{10} = \frac{2}{5}$

$$P((A \cup B)') = P(C) = \frac{2}{10} = \frac{1}{5}$$

$$P(A' \cap B') = P(C) = \frac{2}{10} = \frac{1}{5}$$



Axioms of Probability

> Probability is bounded between 0 and 1.

$$0 \leq P(A) \leq 1$$

Note: "Percent" literally means per one hundred

> Probability of the Sample Space = 1.

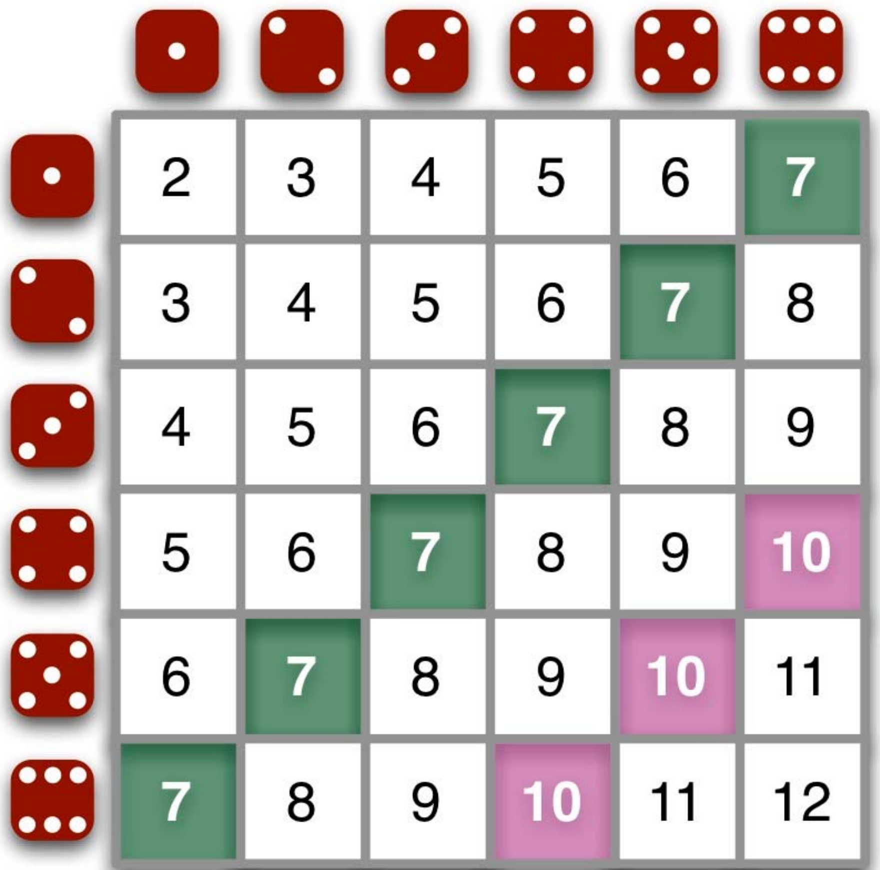
$$P(S) = 1$$

> The probability of finite *mutually exclusive* unions is the sum of their probabilities.

$$P(A \cup B) = P(A) + P(B)$$ If A and B are M.E.

Probability Examples

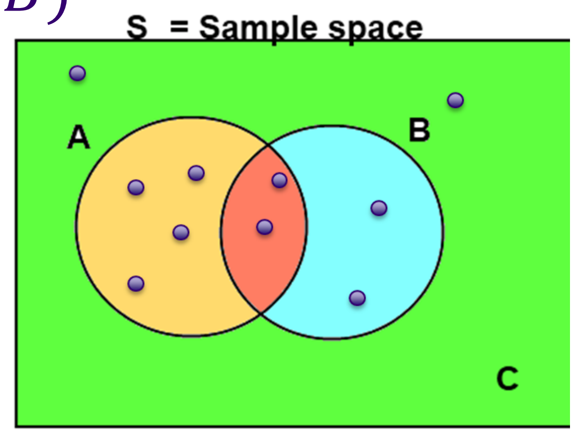
- Probability of rolling a sum of 10?



Mutually Exclusive Events

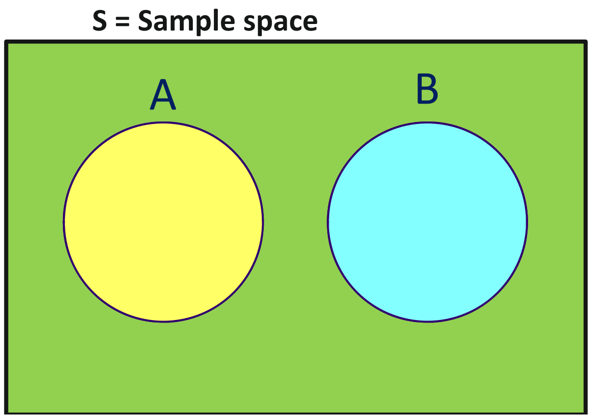
> In all cases, the probability of the union of A and B takes the form:

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



> If A and B are mutually exclusive that means that

$$P(A \cap B) = 0$$
$$P(A \cup B) = P(A) + P(B)$$



Conditional Probability

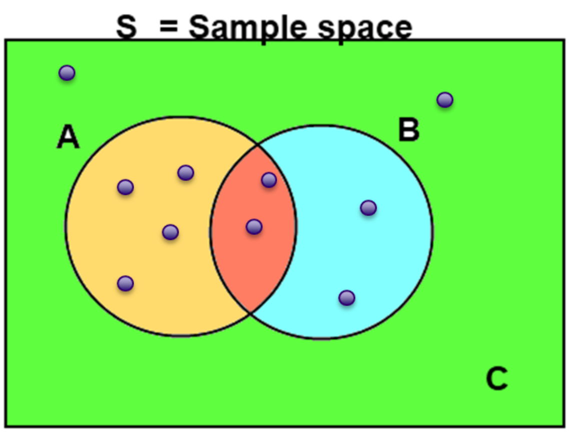
- The probability of A *given* B is written:

$$P(A|B)$$

- And is equal to:

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

$$P(A|B) = \frac{2/10}{4/10} = \frac{2}{4} = \frac{1}{2}$$



Independent Events

- Events A is independent of B if and only if:

$$P(A|B) = P(A)$$

- A being independent of B does NOT imply B is independent of A.

$$P(A|B) = P(A) \quad \not\Rightarrow \quad P(B|A) = P(B)$$

$$P(A|B) = P(A) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(B)P(A) = P(A \cap B)$$

E.g. The event that my boss takes vacation has an impact on when I take vacation, but when I take vacation has no impact on when my boss takes vacation. (i.e., his vacation is independent of mine, but not vice versa)

Independence vs. Mutually Exclusive

- These are not related AT ALL and in fact, are nearly opposite ideas.
- If A is M.E. of B then: $P(A|B) = 0$

B occurring has a HUGE impact on P(A)

- If A is independent of B then: $P(A|B) = P(A)$

Example: The probability the sidewalk is wet given it is raining is very high,
But the probability that it is raining given the sidewalk is wet is lower (if I run my sprinklers often).

Odds

- Odds are expressed as (Count in event favor):(Count not in event favor)
  - Make sure you reduce the fraction first

$$P(A) = \frac{n}{m} = \frac{n}{\underbrace{n}_{\text{Count in favor of A}} + \underbrace{(m-n)}_{\text{Count not in favor of A}}}$$

  - Implies the odds are:

$$n:(m-n)$$

Examples:

If P(A)=5/6, then the odds are 5:1. ‘Five to one’.

If the odds are 3:20, then P(A)=3/23

A straight up sports bet in Vegas has odds 1:1 (50%), but pays 0.95Xbet.

R Demo

Monty Hall Problem

> Famous conditional probability problem that divided statisticians when it came out.

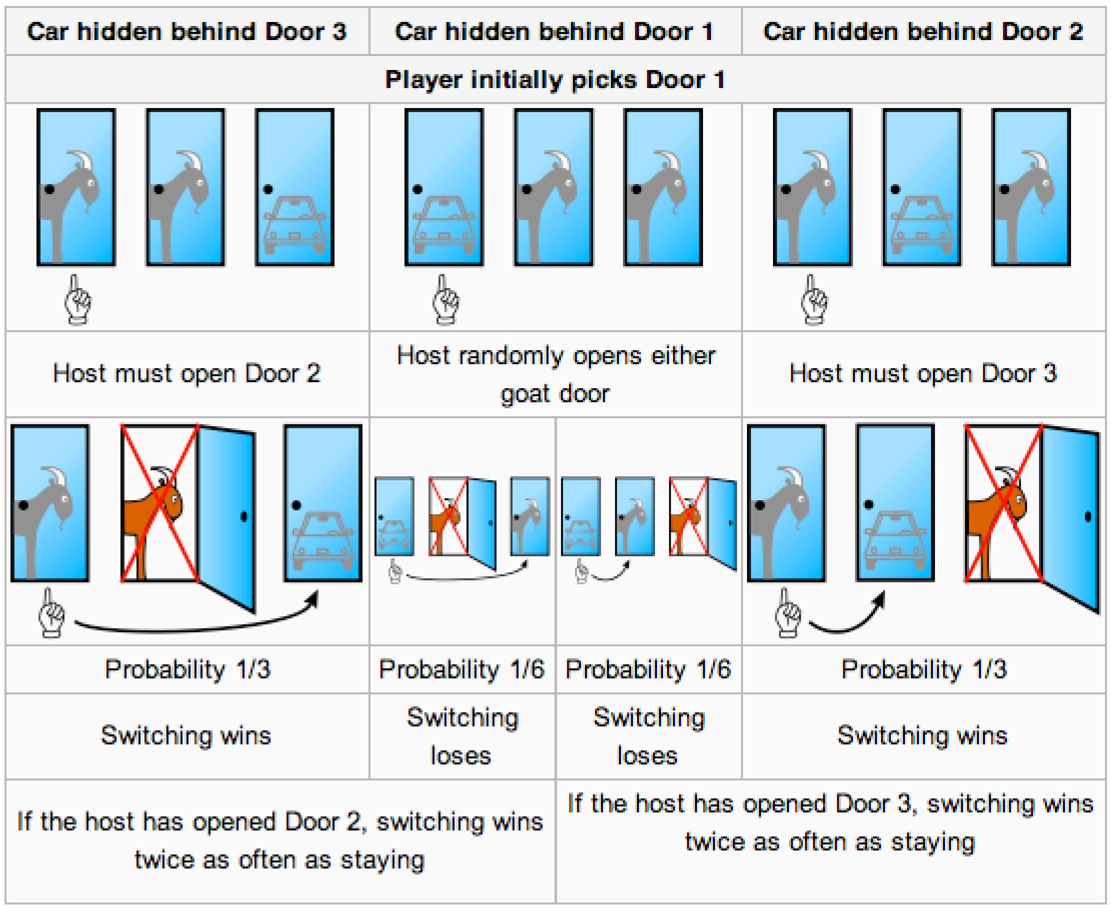
- Start with 3 doors. One prize behind unknown door. Pick a door. Host reveals a separate door with no prize. Then contestant can switch. Should they?

Monty Hall Problem

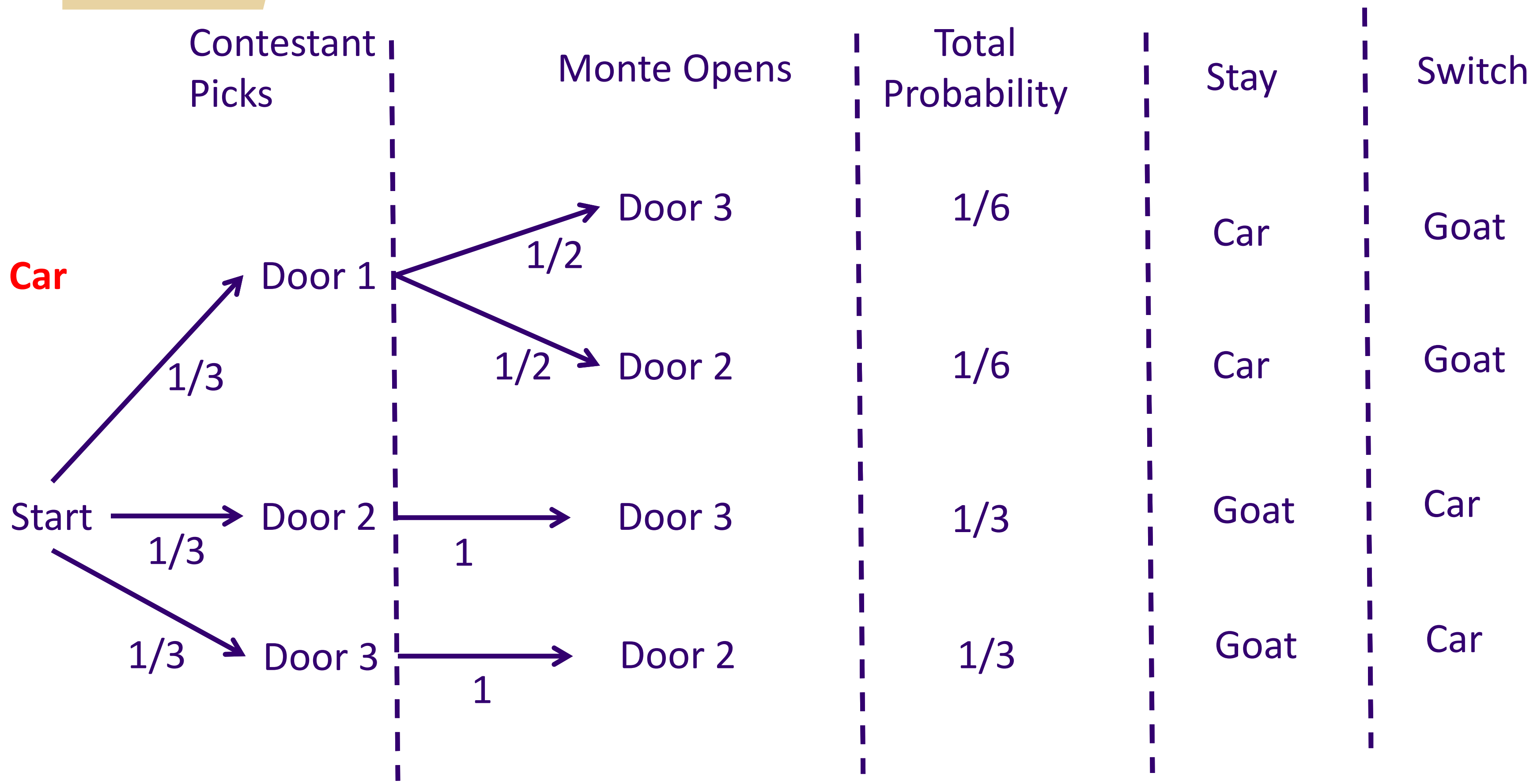
> Start with 3 doors. One prize behind unknown door. Pick a door. Host reveals a separate door with no prize. Then contestant can switch. Should they?

Monty Hall Problem

> Start with 3 doors. One prize behind unknown door. Pick a door. Host reveals a separate door with no prize. Then contestant can switch. Should they?



Monty Hall Problem: Conditional Probabilities



P(Car | Switch) = 2/3
P(Goat | Switch) = 1/3

Monty Hall Problem

- http://www.stayorswitch.com/
- https://en.wikipedia.org/wiki/Monty_Hall_problem

Statistics Review

> Familiar Concepts:
  - Discrete vs. Continuous Distributions
  - Probability
  - Statistics
  - y = mx + b vs $\bar{Y} = \mathbf{M} \cdot \bar{X} + \mathbf{B}$

> These concepts are the focus of this course.

Counting Review

- Factorials
  - Count # ways to order N things = N!
- Permutations
  - Count # of ways to **order** R things from N things = N!/(N-R)!
  - Ordering matters
  - P(N,R)
- Combinations
  - Count # of ways to **group** R things from N things = N!/(R!(N-R!))
  - Ordering doesn't matter
  - C(N,R) or $\binom{N}{R}$
- We will talk about this in depth next class.

Data Distributions (Discrete)

- Discrete Distribution Properties
  - Sum of probability of all possible events must equal 1.
  - Probability of event equal to value of distribution at point.
  - All values strictly in range 0-1.

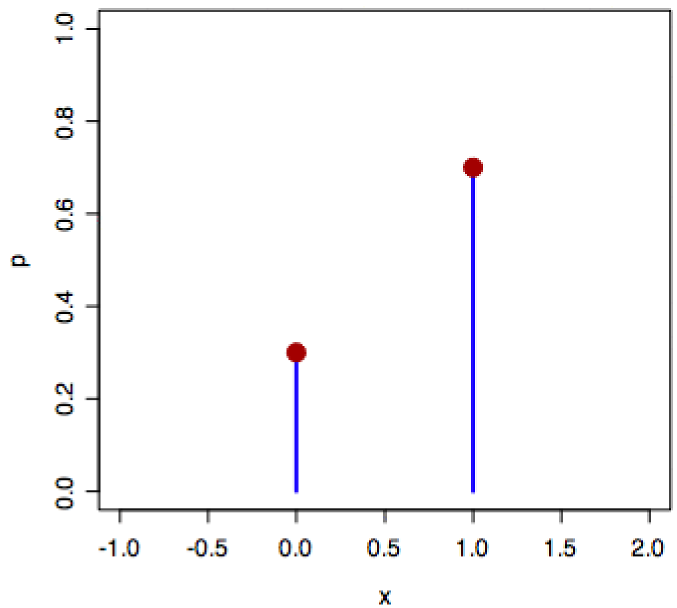
Data Distributions (Discrete)

> Bernoulli (1 event, e.g.: coin flip)

$$P(x) = \begin{cases} p \; if \; x = 1 \\ (1-p) \; if \; x = 0 \end{cases}$$

$$P(x) = p^x (1-p)^{(1-x)} \quad x \in \{0,1\}$$

- Mean = p
- Variance = p(1-p)



Data Distributions (Discrete)

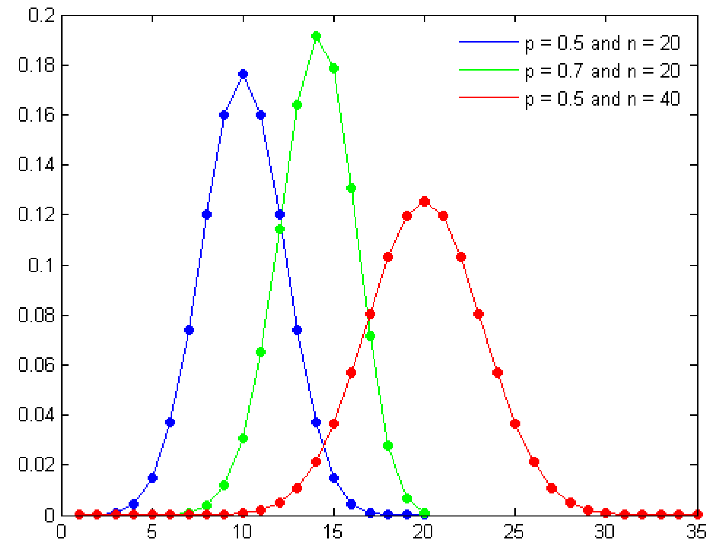
> Binomial (Multiple Bernoulli's Events)

– Multiple Independent events = Product of Bernoulli Probabilities

$$P(x|N,p) = \binom{N}{x} p^x (1-p)^{(N-x)}$$

– Mean = np

– Variance = np(1-p)



Note: for larger n, we approximate this by a normal distribution.

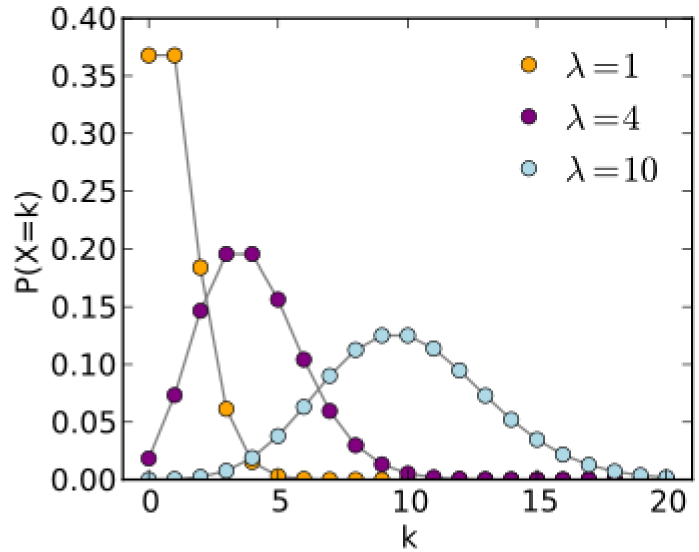
Data Distributions (Discrete)

> Poisson (Count of number of events in a time span)

$$P(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- Mean = λ
- Variance = λ

Interpret as the rate of occurrence of an event is equal to lambda in a finite period of time.



R Demo

Discrete distributions

Data Distributions (Continuous)

> Continuous Distribution Properties

- Area under the curve must be equal to 1.
- Probability a range of values of an event equal to AREA under curve.
- No negative values.
- Probability of a single, exact value is 0.

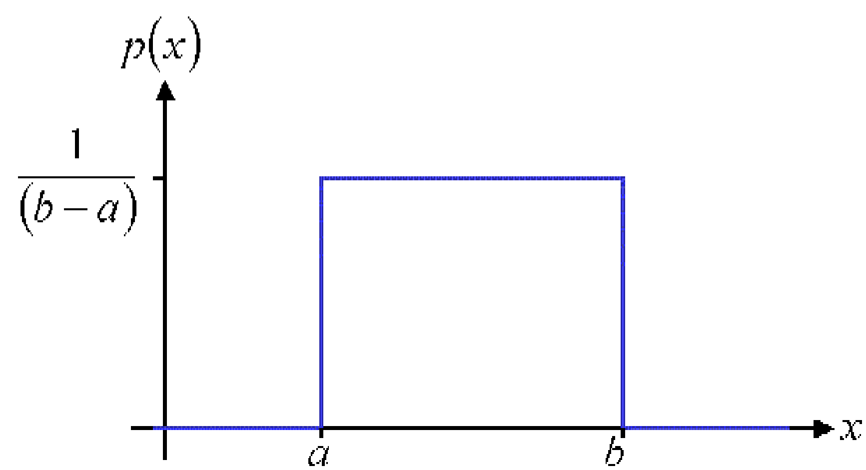
Data Distributions (Continuous)

> Uniform (flat, bounded)

$$P(x) = \begin{cases} \dfrac{1}{(b-a)} \; if \; a \leq x \leq b \\ 0 \; if \; x < a \; or \; x > b \end{cases}$$

> Used for parameter priors. (future discussion)

- Mean=(a+b)/2
- Variance=(1/12)(b-a)^2

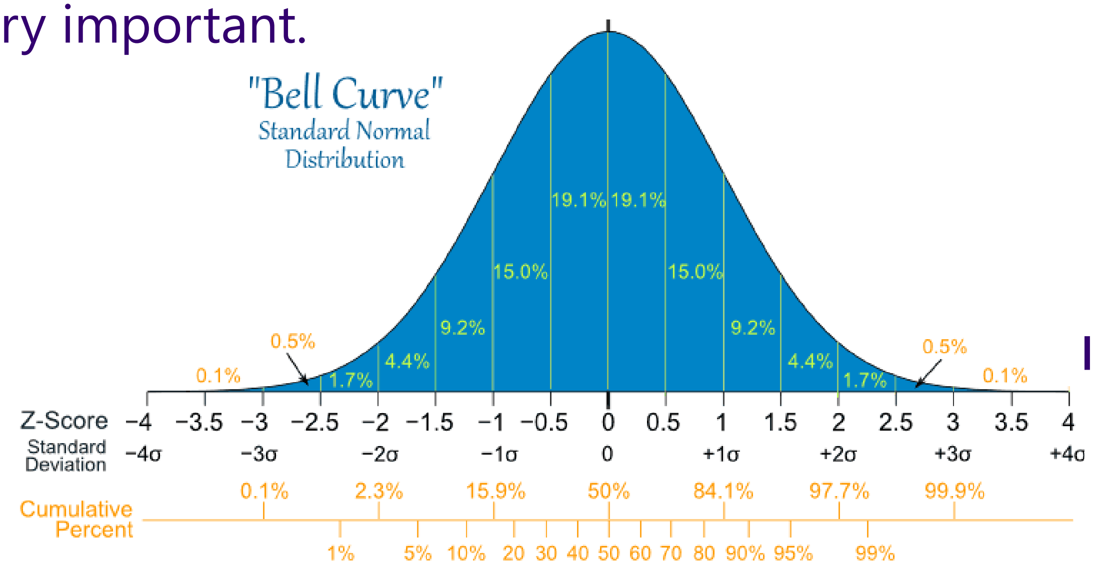


Data Distributions (Continuous)

> Normal (Gaussian) distribution
  - Most common and occurs naturally.
  - Defined by a mean and variance only. (standard = N(0,1))

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

  - Has very nice properties.
  - Tests for normality are very important.



Data Distributions (Continuous)

- Student's T (normal for small samples)
  - Important for hypothesis testing smaller sample sizes.
  - Used for:
    - Testing of mean value when st. dev. is unknown.
    - Testing difference between two distribution means.
  - Looks very similar to the normal distribution.

R Demo

Continuous distributions

Simulations

> Used for complex distributions
> Can test distributional assumptions
> Simulate a conditional probability hierarchy
> Large number of realizations
> Use system.time() from base or microbenchmark() from microbenchmark package.
> R Demo

Testing Statistical Software

- Usual test processes apply: Need to build test cases
- Test cases must be repeatable (e.g. set.seed())
- Build test cases as you go: Test driven development

R DEMO

R review and summary statistics

> Purpose: To gain a clear understanding of your data.
  - How large is it?
  - What columns are of interest?
  - Missing data?
  - Outliers?