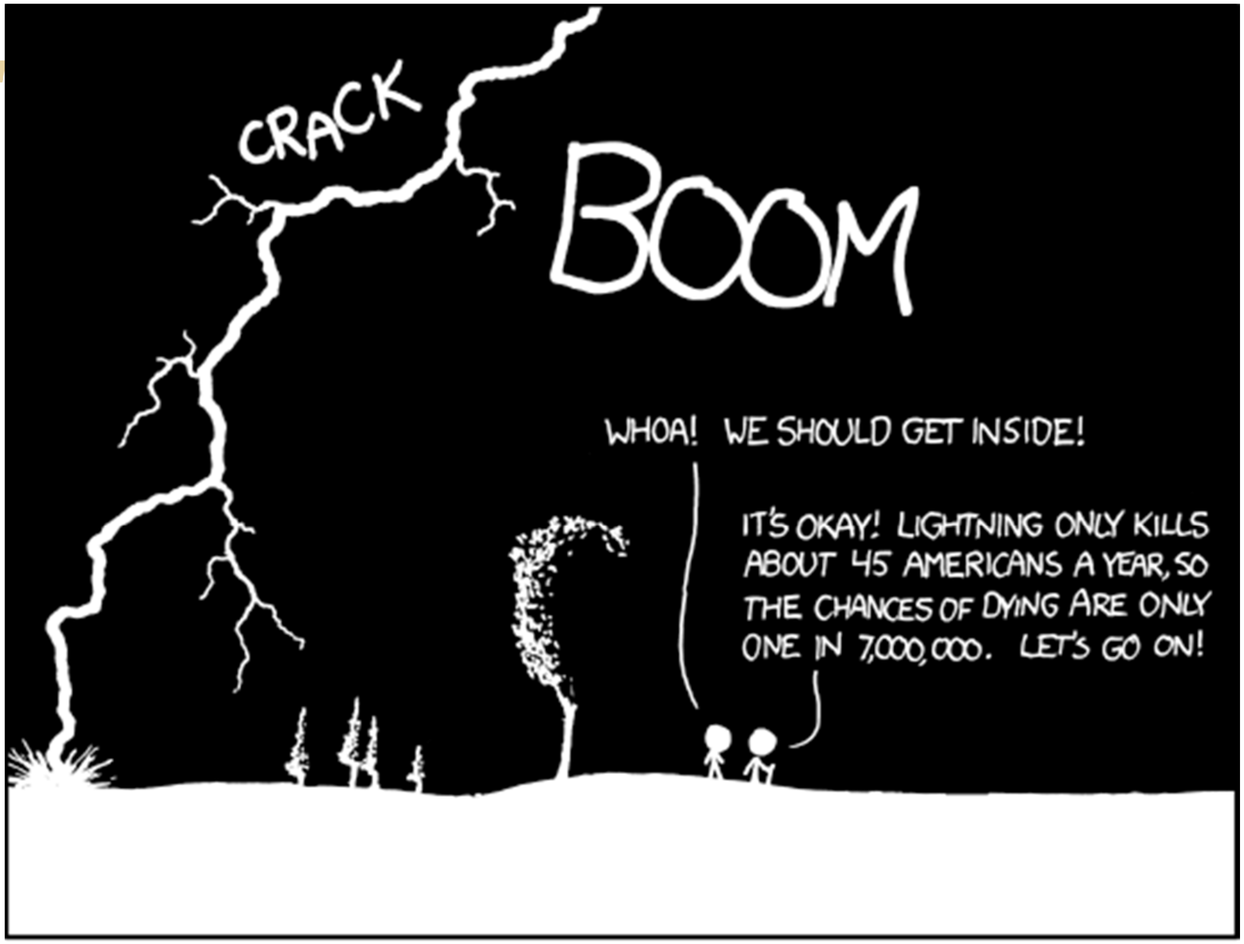
UNIVERSITY *of* WASHINGTON

# 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
- > Loading data

# Review

- > Distributions
  - – Discrete: Bernoulli, Binomial, Poisson
  - – Continuous: Uniform, Normal, Student’s T
- > Numerical and Visual Exploration of Data
- > Transformations
- > Simpson’s Paradox

# Review

## Summary statistics

> Sample mean = $\mu$ = sum($x_i$)/n
> sample var = $\sigma$ = sum(($\mu$ – $x_i$)^2) / (n – 1)
> Sample std = sqrt($\sigma$)
> Standard error of the sample mean = se = std/ sqrt(n)

# 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 |
|---|---|---|---|---|---|---|---|---|
| ↑ | | ↑ | | ↑ | | ↑ | | ↑ |
| 1st book | | 2nd book | | 3rd book | | 4th book | | 5th 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?

| | | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 |
| | 3 | 4 | 5 | 6 | 7 | 8 |
| | 4 | 5 | 6 | 7 | 8 | 9 |
| | 5 | 6 | 7 | 8 | 9 | 10 |
| | 6 | 7 | 8 | 9 | 10 | 11 |
| | 7 | 8 | 9 | 10 | 11 | 12 |

# Counting in R

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

# Probability

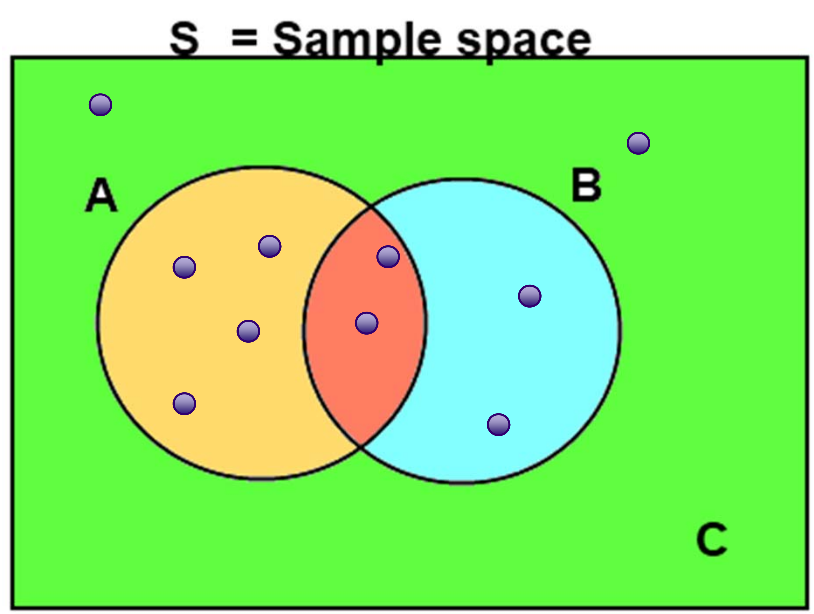
> The Probability of an event, A, is the number of ways A can occur, divided by the number of total possible outcomes in our Sample Space, S.

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

> If • is an event, then

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

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

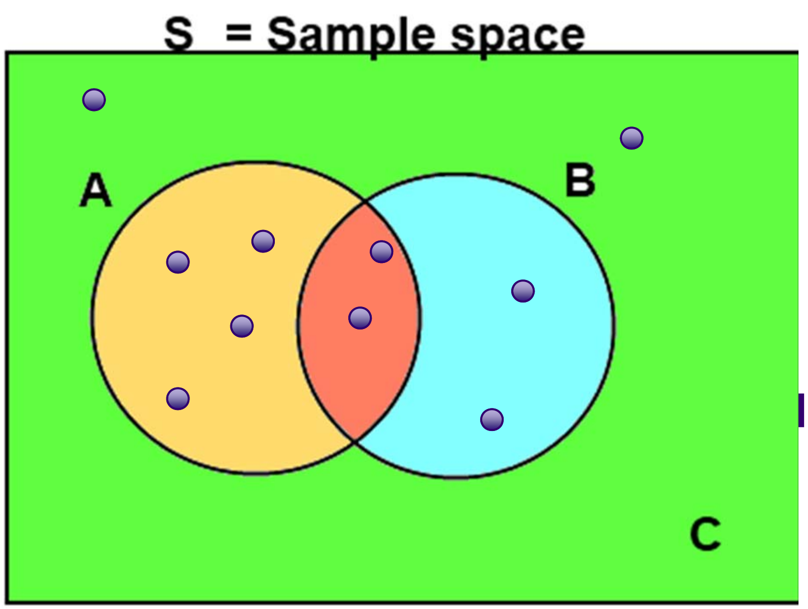


# Probability

- 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.

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

# Probability Examples

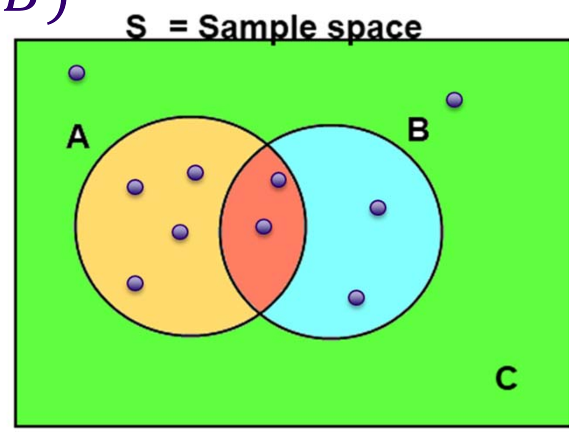
> Probability of rolling a sum of 10?

| | | | | | |
|---|---|---|---|---|---|
| 2 | 3 | 4 | 5 | 6 | 7 |
| 3 | 4 | 5 | 6 | 7 | 8 |
| 4 | 5 | 6 | 7 | 8 | 9 |
| 5 | 6 | 7 | 8 | 9 | 10 |
| 6 | 7 | 8 | 9 | 10 | 11 |
| 7 | 8 | 9 | 10 | 11 | 12 |

# Mutually Exclusive Events

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

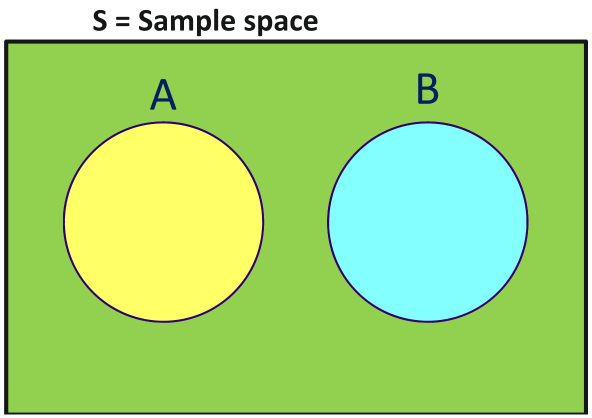
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$




> 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

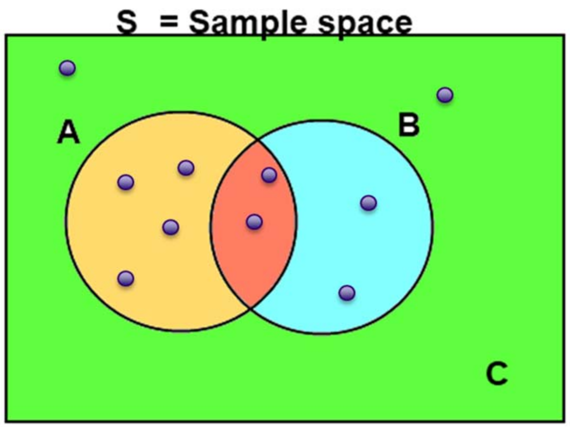
- The probability of A *given* B is written:

$$P(A|B)$$

- 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

$$P(A) = \frac{n}{m} = \frac{n}{n + (m - n)}$$

Count in favor of A

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

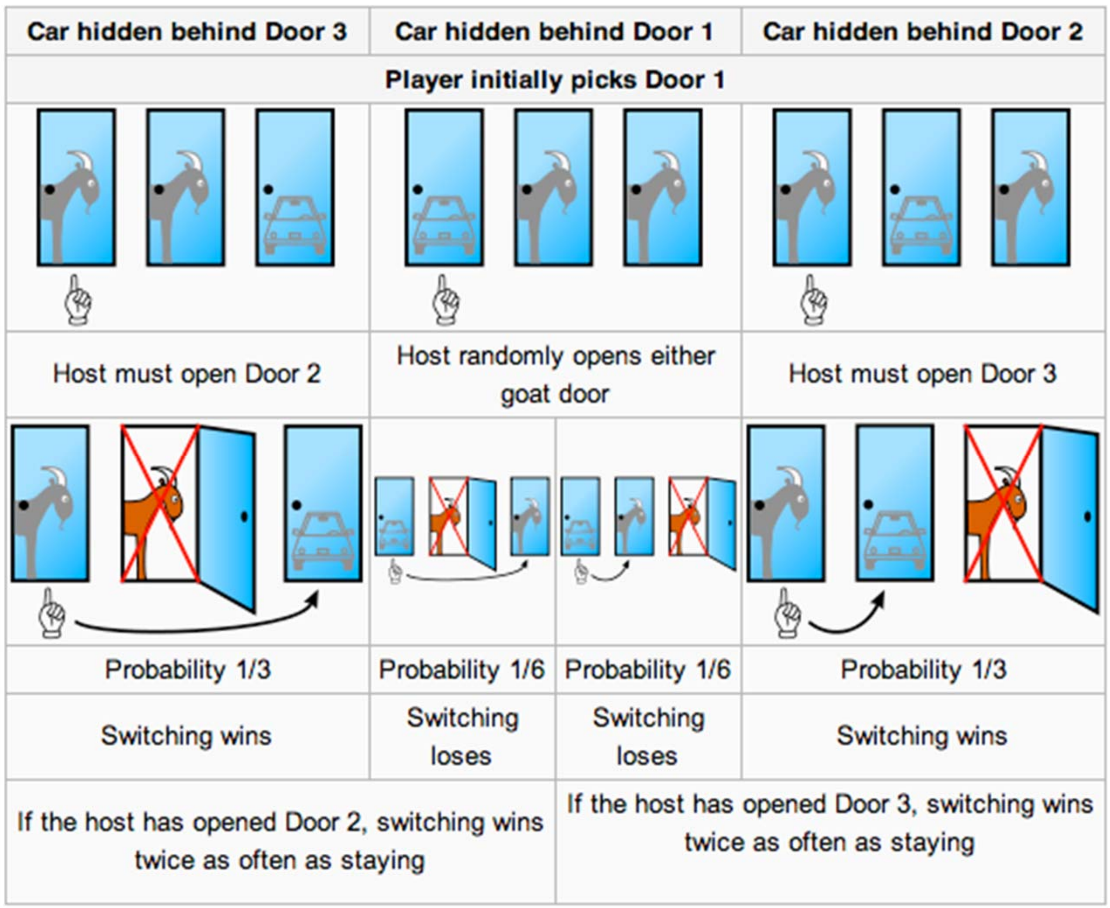
- Famous conditional probability problem that divided statisticians when it came out.
  - Start with 3 doors. One prize behind unknown door. Pick a door. Host reveals a separate door with no prize. Then contestant can switch. Should they?

# Monty Hall Problem

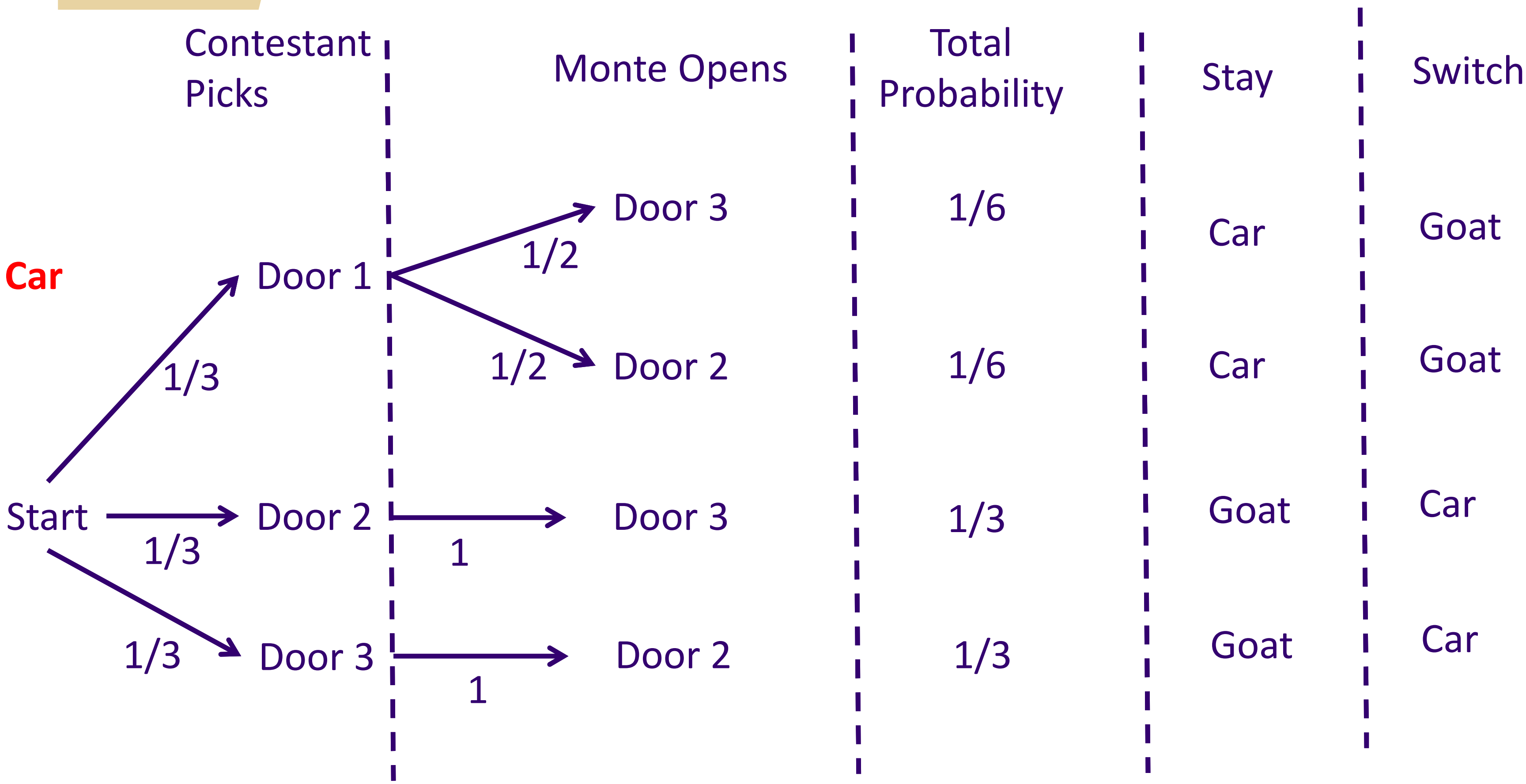
> Start with 3 doors. One prize behind unknown door. Pick a door. Host reveals a separate door with no prize. Then contestant can switch. Should they?

# 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

# 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

# Dealing with Missing Data

- Reasons for missing data
  - Recording failure (mechanical/software failures)
  - Reporting failure (human decisions)
  - Translation failure (data transferring/parsing errors)
  - Inconsistent recording proceedures
- Outliers may also be treated as missing data.

# Dealing with Missing Data

| Type | Benefits | Disadvantages | Notes |
|---|---|---|---|
| Drop Missing | -Speed | -Data Loss | |
| Substitution | - Speed | - Bias | |
| Mean/Median/Mode Fill | -No Data Loss | -Variance Reduction | |
| X~F(independents) | -More Accurate<br>-No Data Loss | -Slower | -Needs most columns to be filled out<br>-Hard on ind. data |
| knn | -More Accurate<br>-No Data Loss | -Slower<br>-Dependent on distance function<br>- Bias | Forward fill<br>Backward fill |
| X~F(y,independents) | -Very accurate<br>-No Data Loss | -Slower<br>-Need y | -Only on training set! |

# Dealing with Missing Data: Using Outside or New Data Sources

> Don't forget to explore outside or new data sources to help fill-in missing data.
> With the advent of free public data and bigger data sources, this is gaining popularity as a tool for imputation.
> Unstructured text is a major source of data.
> Ex:
  - Caesar's uses public reviews on websites to mine for customer sentiment about hotel rooms.
  - Zillow uses text descriptions of properties to fill in missing data about # bedrooms, # bathrooms, sq. footage, and various amenities.
  - Subject to human stupidity.

Yelp Rating for Circus-Circus: 2/5
Text Description: "My son and I stayed here. The service was great, the room was great, but it turns out my son is deathly afraid of clowns."

# Dealing with Missing Data: Variance and Multiple Imputation

- Imputation tries and maintain the intrinsic variance in the data set.
- Multiple different predictions are made for each missing data point. (Using previous methods)
- Hypothesis testing and predictions are made on all imputed sets to gauge the variance in the outcomes.
- R package ‘Amelia’ does this and creates a nested list of data frames.
- Lots of details at: https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf
- R demo

# Getting Data

- Files
  - Csv: read.csv
  - Txt: read.table
- Web/HTML
  - readLines
  - XML, xpath
  - http://gastonsanchez.com/work/webdata/getting_web_data_r4_parsing_xml_html.pdf
- Databases
  - Sqlite: sqldf, RSQLite packages
    - Sqlite example
  - MongoDB: rmongodb package
  - Postgresql: RPostgreSQL package

# Storing Data

- .csv – write.csv()
- .txt – write.txt()
- .Rdata – save()
  - Workspaces are very compressed compared to csv
- Databases
  - Sqlite: sqldf, RSQLite packages
    - Sqlite example
  - MongoDB: rmongodb package
  - Postgresql: RPostgreSQL package

# SQL and R

- Handle datasets larger than memory
- Support several common databases.
- e.g. SQLite: http://www.r-bloggers.com/r-and-sqlite-part-1/
- Or with dplyr: https://cran.r-project.org/web/packages/dplyr/vignettes/databases.html
- And, many other packages and references: search around.
- R Demo

# Assignment

- Homework 2:
  - Write an R-script to compute the Monty Hall Probabilities with simulations (get probabilities AND variances for switching and not switching).
  - You should submit:
    - **ONE R-script** that outputs the probabilities and variances.
    - Submission should include a proof of correct answer: chart and table.
  - Read Intro to Data Science Chapter 7 and 10.
  - Read Statistical Thinking for Programmers Ch. 4.