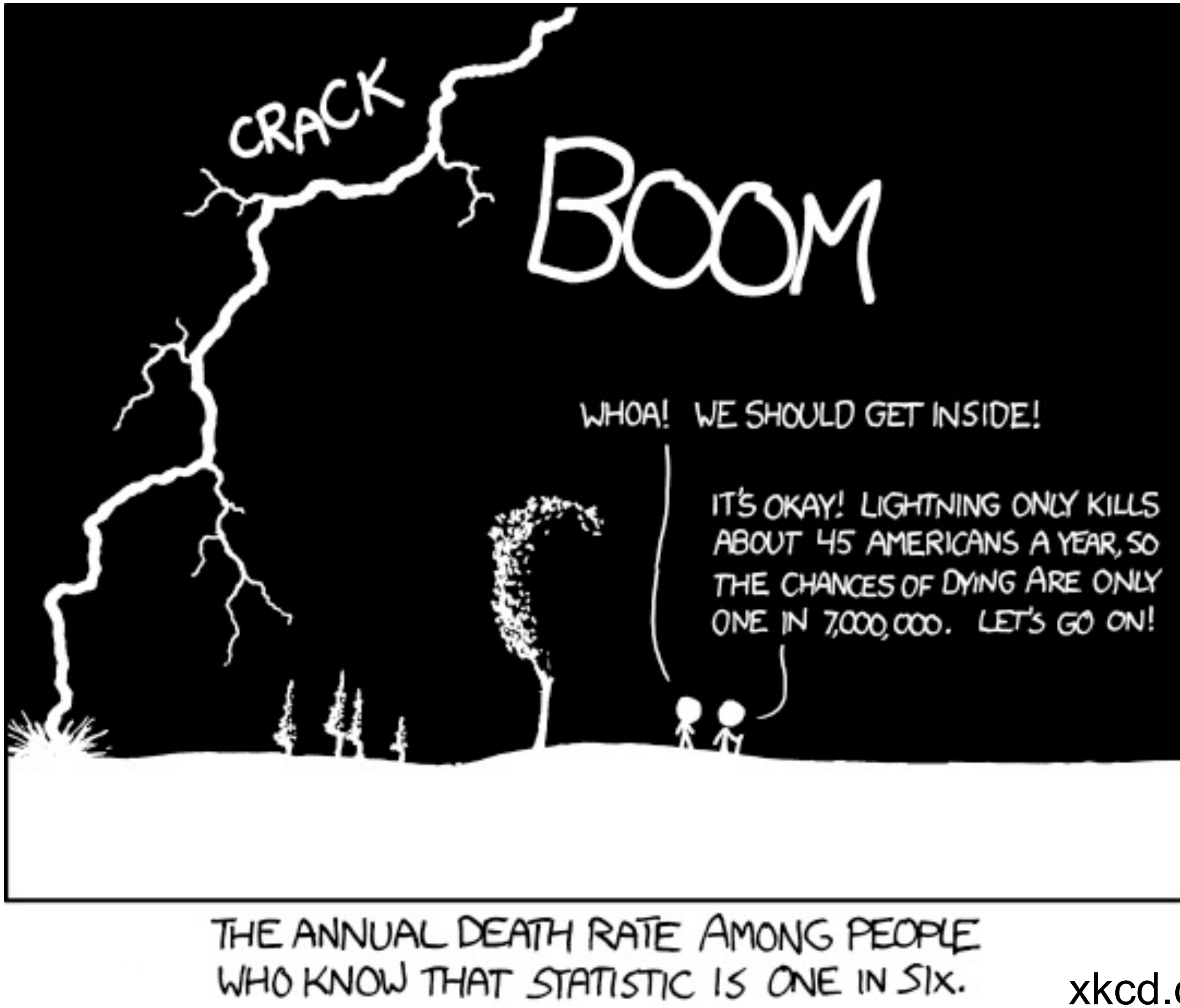
# Probability Review




xkcd.com/795/

Zach Wood-Doughty CS 396 Winter 2022

Some slides borrowed from Bryan Pardo and Elizabeth Tipton lecture2demo.py

Account

Dashboard

Courses

Calendar

History

Help

```
import argparse
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf


def observed(n=100, c_dim=6, ols="y ~ a"):
    c = np.random.randint(1, 1 + c_dim, n)
    a = np.random.binomial(n=1 + c_dim - c, p=0.5, size=n)
    a = (a > 0).astype(np.int32)
    y = np.random.binomial(n=a + c, p=0.5)
    df = pd.DataFrame(data=dict(c=c, a=a, y=y))
    return smf.ols(ols, data=df).fit().params['a']


def randomized(n=100, c_dim=6, ols="y ~ a"):
    c = np.random.randint(1, 1 + c_dim, n)
    a = np.random.binomial(n=1, p=0.5, size=n)
    y = np.random.binomial(n=a + c, p=0.5)
    df = pd.DataFrame(data=dict(c=c, a=a, y=y))
    return smf.ols(ols, data=df).fit().params['a']
```

https://canvas.northwestern.edu/courses/157017/files/folder/Code%20and%20Data?

# Tentative schedule: Weeks 1-3

1. Motivating causal inference
   - Simpson's paradox
   - Counterfactuals
   - Randomized experiments
2. Review of fundamentals
   - Probability and statistics
   - Graphical models and conditional independence
   - Connecting potential outcomes to observational data
   - **HW1 Out**
3. Basic methods in causal inference (no class Monday)
   - Simple confounding and identification
   - **HW1 Due**

# Tentative schedule: Weeks 4-6

4. Estimators of causal effects
   - Outcome models
   - Propensity score models
5. Unmeasured confounding and identification
   - **Project proposal due**
6. Structure learning
   - Testing for conditional independences
   - PC and GES Algorithms

# Tentative schedule: Weeks 7-10

7. Missing data
   – **Project update due**
8. Measurement error and proxies
9. Selection bias and case-control studies
   – **Project presentations**
10. Additional topics
    – **Presentation peer feedback due**
11. Project report due on Monday, Mar 14

# Final projects

1. Pick a dataset (and a group)
   - What is the causal question you're interested in?
   - Does this dataset contain enough data to answer it?
2. Pick a graphical model that describes the data
   - Use domain knowledge and data-driven methods
3. Identification
   - Theoretical justification for how to infer causality
4. Estimation
   - Use a method to compute the effect from data
5. Additional methods: missing data, selection bias, etc.
6. Analysis of the results

# Axioms of Probability

- Let there be a space S composed of a countable number of events

$$S \equiv \{e_1, e_2, e_3, \dots. e_n\}$$

- The probability of each event is between 0 and 1

$$0 \leq P(e_1) \leq 1$$

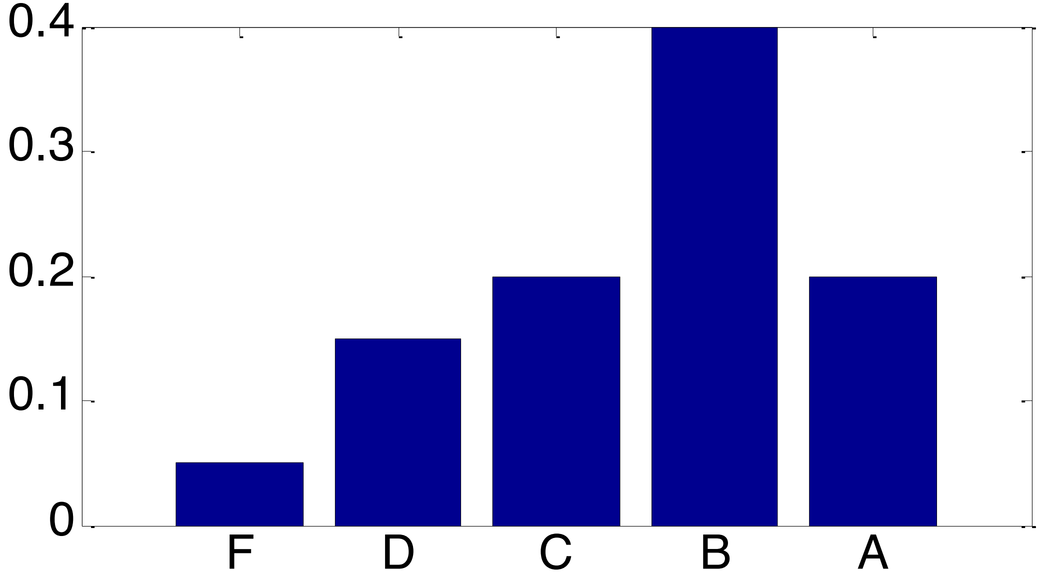
- The probability of the whole sample space is 1

$$P(S) = 1$$

- **When two events are mutually exclusive**, their probabilities are additive

$$P(e_1 \vee e_2) = P(e_1) + P(e_2)$$

# Discrete Random Variables




| Grade | Probability |
|---|---|
| A | 0.2 |
| B | 0.4 |
| C | 0.2 |
| D | 0.15 |
| F | 0.05 |

- P(Grade) is a distribution over possible grades
- Each grade is mutually exclusive
- Probabilities sum to 1

# Boolean Random Variable

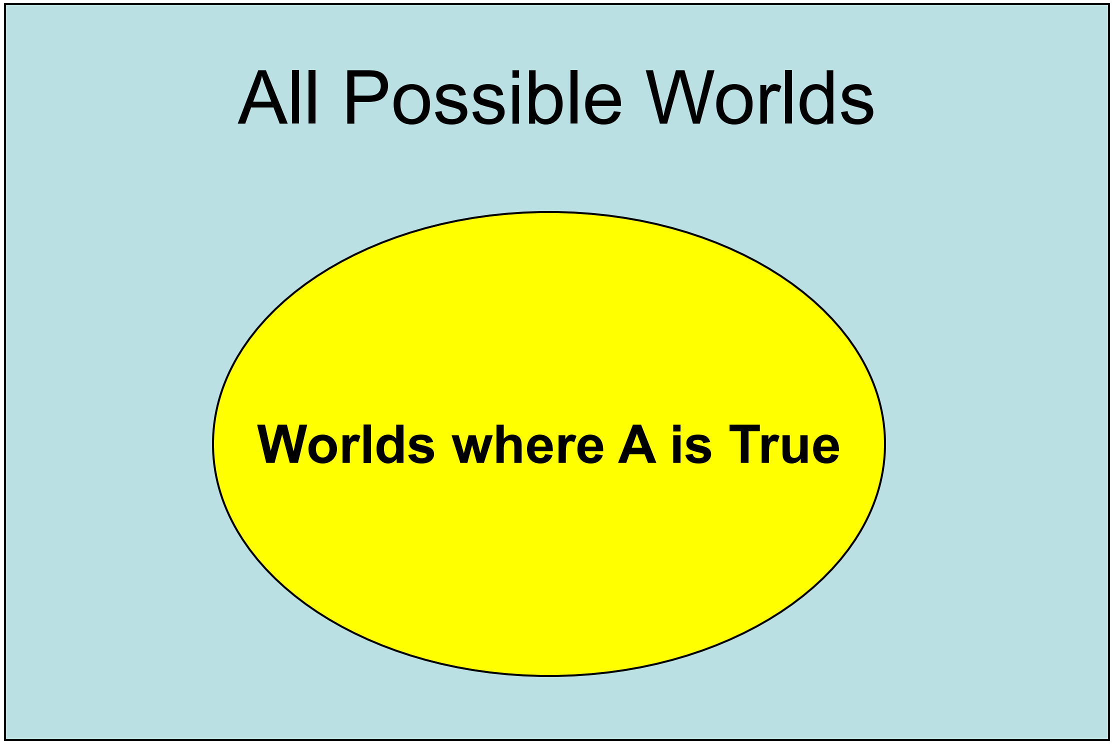
- Boolean random variable: A random variable that has only two possible outcomes

  e.g.

  **X** = “Tomorrow’s high temperature > 60” has only two possible outcomes

  As a notational convention, **P(X)** for a Boolean variable will mean **P(X=“true”)**, since it is easy to infer the rest of the distribution.

# Vizualizing P(A) for a Boolean variable

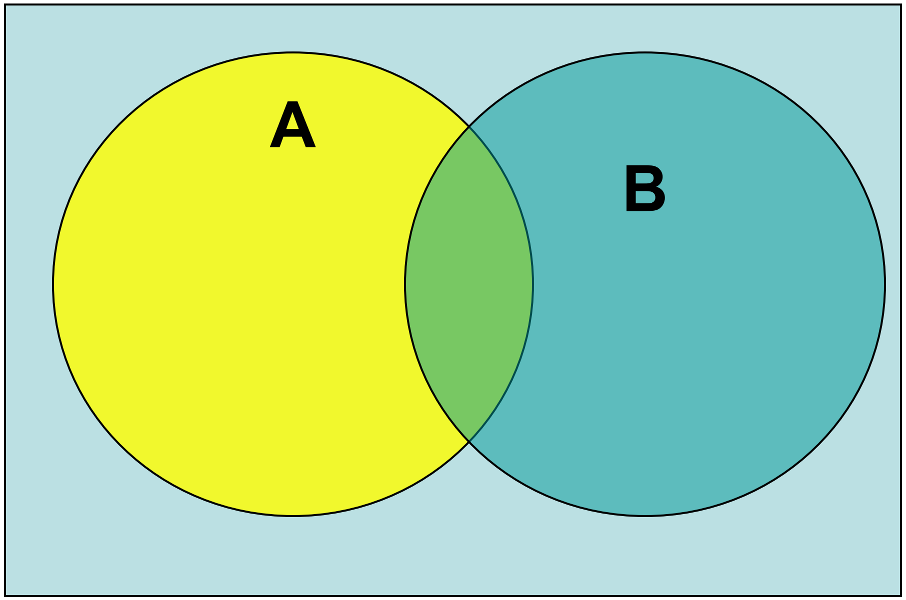



$0 \le P(A) \le 1$

If a value is over 1 or under 0, it isn't a probability

$$P(A) = \frac{\text{area of yellow oval}}{\text{area of blue rectangle}}$$

# Visualizing two Booleans




$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

# Independence

- variables A and B are said to be *independent* iff…

$$P(A)P(B) = P(A \wedge B)$$

# Bayes Rule

- Definition of Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

- Corollary:

  The Chain Rule

$$P(A \mid B)P(B) = P(A \wedge B)$$

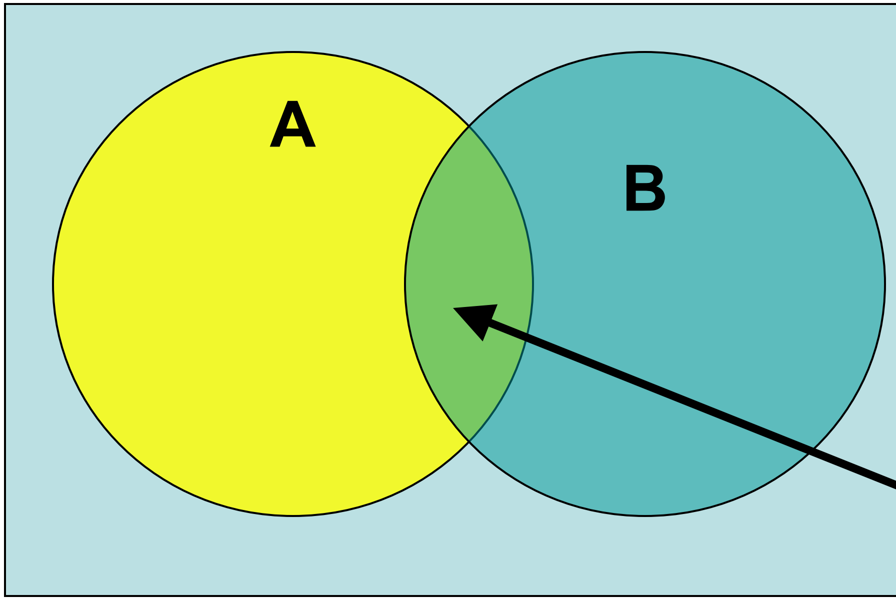
- Bayes Rule

  (Thomas Bayes, 1763)

$$P(B \mid A) = \frac{P(A \wedge B)}{P(A)}$$

$$= \frac{P(A \mid B)P(B)}{P(A)}$$

# Conditional Probability




The conditional probability of A given B is represented by the following formula

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

**Overlap implies NOT independent**

Can we do the following?

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(A)P(B)}{P(B)}$$

Only if A and B are ***independent***

# The Joint Distribution

- Truth table lists all combinations of variable assignments
- Assign a probability to each row
- Probabilities sum to 1

| A | B | C | Prob |
|---|---|---|---|
| 0 | 0 | 0 | 0.1 |
| 0 | 0 | 1 | 0.2 |
| 0 | 1 | 0 | 0.1 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.2 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.05 |

# Using The Joint Distribution

- Find P(A)
- Sum the probabilities of all rows where A=1

$$P(A) = 0.05 + 0.2 + 0.25 + 0.05 = 0.55$$

| A | B | C | Prob |
|---|---|---|---|
| 0 | 0 | 0 | 0.1 |
| 0 | 0 | 1 | 0.2 |
| 0 | 1 | 0 | 0.1 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.2 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.05 |

# Using The Joint Distribution

- Find P(A|B)

$$p(A \mid B) = \frac{p(A, B)}{p(B)}$$

$$p(B = b) = \sum_{a \in \{0,1\}} p(A = a, B = b)$$

= (0.25+0.05)
÷ (0.25+0.05 + 0.1+0.05)
= 0.3 ÷ 0.45
= 0.667

| **A** | **B** | **C** | **Prob** |
|---|---|---|---|
| 0 | 0 | 0 | 0.1 |
| 0 | 0 | 1 | 0.2 |
| 0 | 1 | 0 | 0.1 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.2 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.05 |

# Using The Joint Distribution

Are A and B Independent?

$P(A, B) = 0.25 + 0.05$

$P(A) = 0.3 + 0.2 + 0.05$

$P(B) = 0.3 + 0.1 + 0.05$

$P(A) \times P(B) = 0.55 \times 0.45$

$P(A, B) = 0.3 \neq 0.248$

A and B NOT independent

| **A** | **B** | **C** | **Prob** |
|---|---|---|---|
| 0 | 0 | 0 | 0.1 |
| 0 | 0 | 1 | 0.2 |
| 0 | 1 | 0 | 0.1 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.2 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.05 |

# Why not use the Joint Distribution?

- Given *m* boolean variables, we need to estimate $2^m$ values.

- 20 yes-no questions = a million values

- How do we get around this combinatorial explosion?
  - Assume independence of variables!

# …back to independence

- My height is independent of my favorite basketball team
- This is **domain** knowledge, typically supplied by the problem designer
- Independence implies:

$$A \perp B \Rightarrow p(A \mid B) = p(A)$$

$$A \perp B \mid C \Rightarrow p(A, B \mid C) = p(A \mid C)p(B \mid C)$$

# Let's show that

assuming independence...

$$P(A \wedge B) = P(A)P(B)$$

plus the chain rule...

$$P(A \wedge B) = P(A \mid B)P(B)$$

imply...

$$P(A)P(B) = P(A \mid B)P(B)$$

which means...

$$P(A \mid B) = P(A)$$

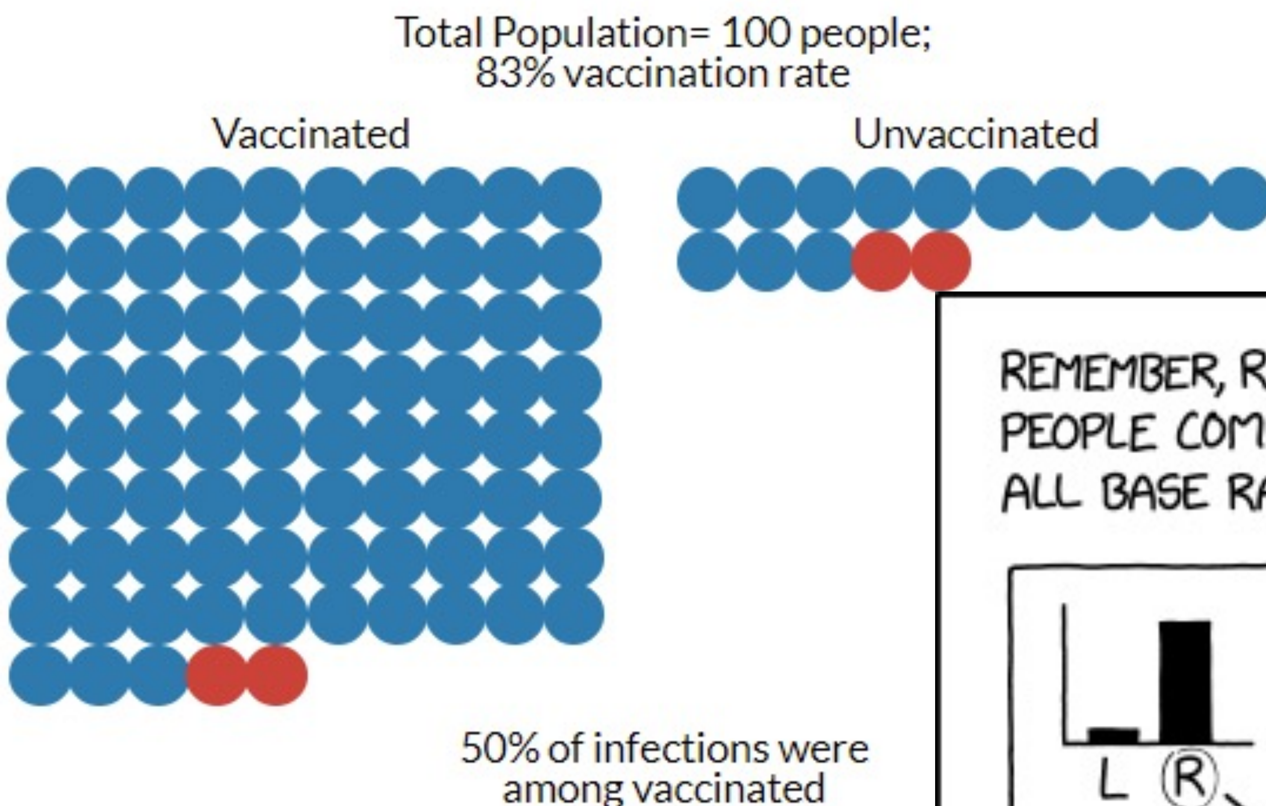
# Reasoning with Probability

P(cancer) = 0.01
P(positive test | cancer) = 0.97
P(positive test | no cancer) = 0.02

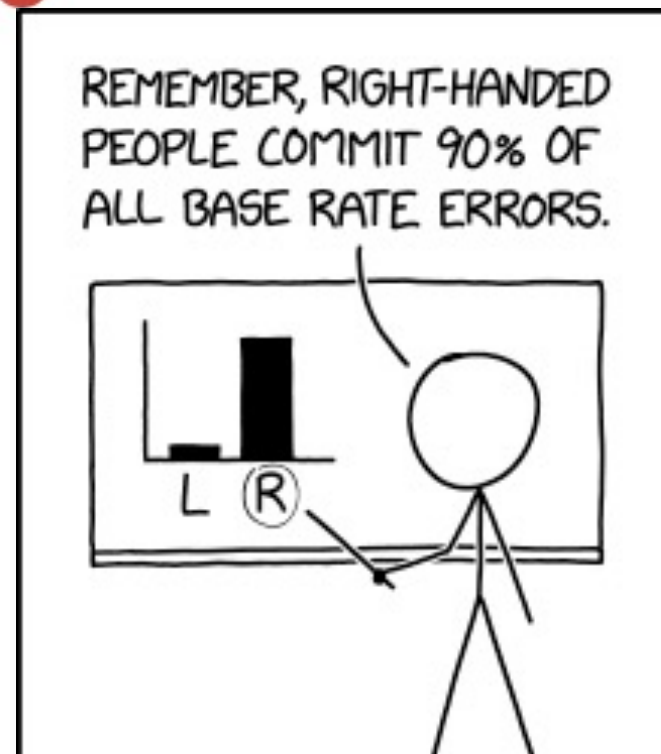
What is p(cancer | positive test)?

| Test | Can | Prob |
|---|---|---|
| 1 | 1 | A |
| 0 | 1 | B |
| 1 | 0 | C |
| 0 | 0 | D |

# Base Rate Fallacy

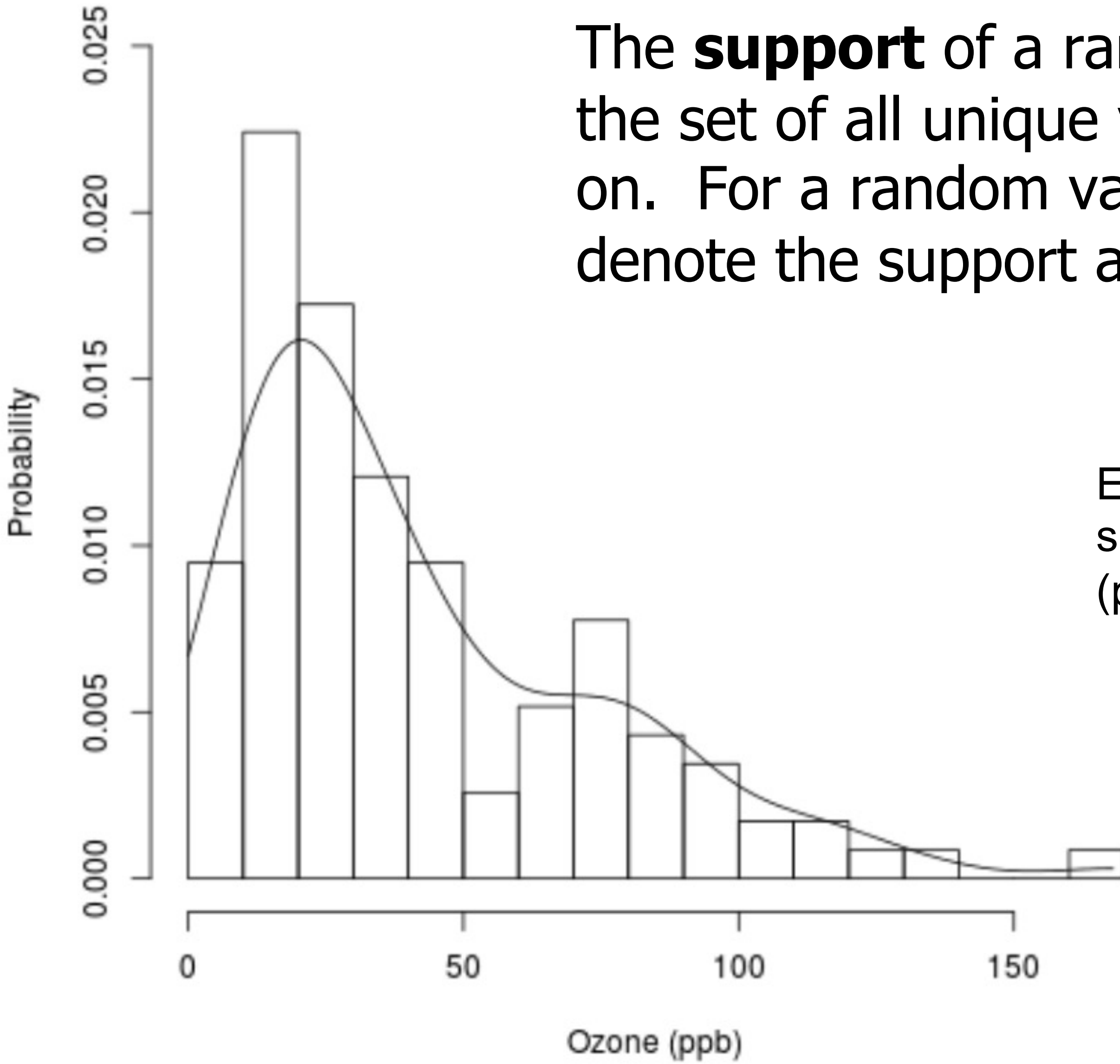



yourlocalepidemiologist.substack.com




xkcd.com/2476/

# Supports




The **support** of a random variable is the set of all unique values it can take on. For a random variable $X$ we denote the support as $S_X$.

Example, the support of Ozone (ppb) is 0 to 155.

# Expectation

The **expectation** of a random variable is its average value

$$E(X) = \mu_X$$

Expectation is determined by the marginal distribution function:

$$E(X) = \sum_{x \in S_X} x \Pr(X = x)$$

**Conditional expectation** looks similar:

$$E(Y|X = x) = \sum_{y \in S_Y} yPr(Y = y|X = x)$$

# Properties of expectation

1. If $T$ is a binary random variable (with $S_T = \{0,1\}$) then $E(T) = \Pr(T = 1)$.

2. If $c$ is a constant (fixed) value, then $E(c) = c$.

3. Expectations are **linear**. For fixed constants $a, b, c,$
$$E(aX + bY + cZ) = aE(X) + bE(Y) + cE(Z)$$

4. If variables X, Y, and Z are mutually independent, then
$$E(XYZ) = E(X)E(Y)E(Z)$$

# Estimators

- Suppose X is a six-sided die.
  What is E[X]? How do we find it?

- $\theta = E(X) = \sum_{x \in S_X} x \Pr(X = x)$

- How do we compute E[X] without knowing Pr(X = x)?

- $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} X_i$

- We call $\hat{\theta}$ an estimator for $\theta =$ E[X]

# Recall: synthetic data example

C: result of a k-sided die roll

Flip 1 + k - C coins. Then

A: 1 if at least one head

0 otherwise

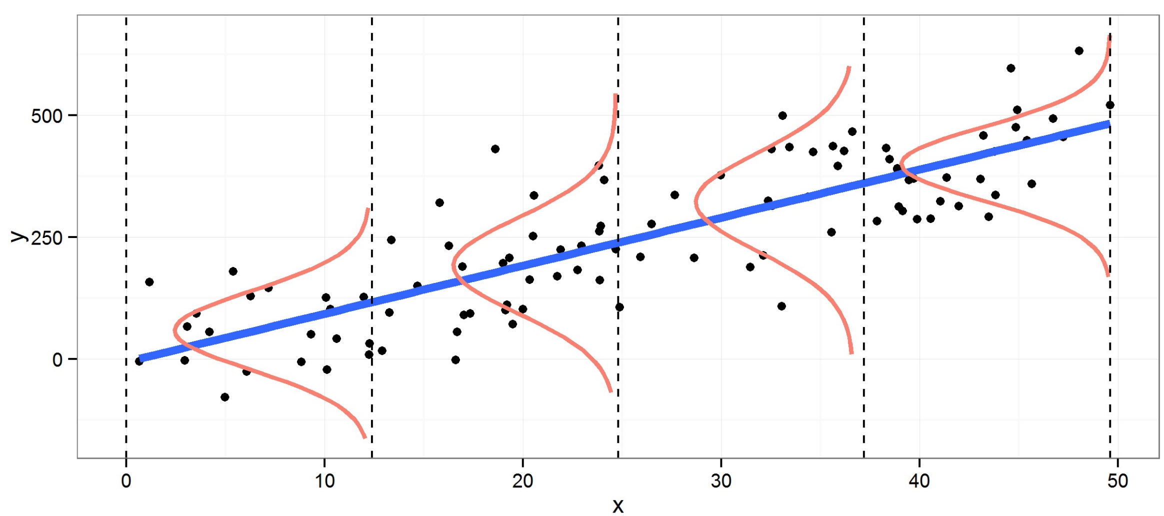
Flip C + A coins. Then Y is the total number of heads

# Prediction versus inference

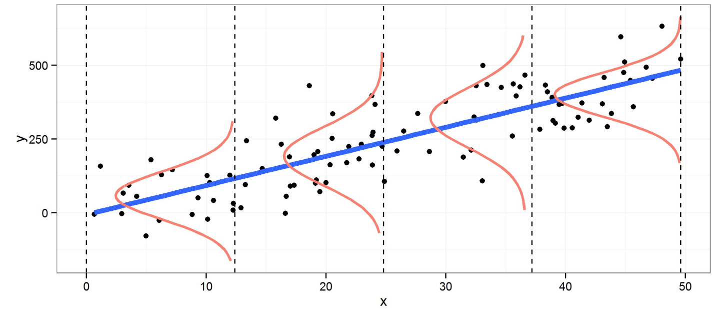
- In machine learning (ML), we might say "I want to know how to predict Y using the C and A."

- In statistics, we might say "I want to know the parameters that define Y's behavior given C and A."

- In both, we'll fit models such as E[Y | A, C]



# Effect of A on Y in synthetic data




When A is randomized,

```
smf.ols(“y ~ a”, data=df).fit().params[“a”]
```

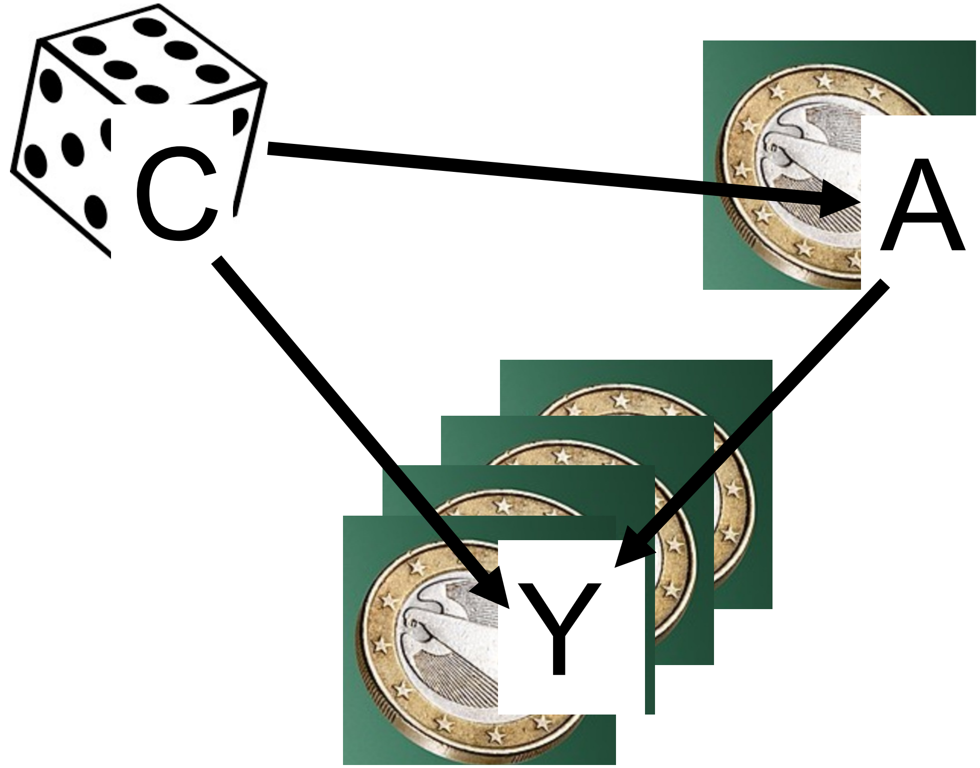
| | |
|---|---|
| $E[Y \mid A=1]$ | $E[Y \mid A=0]$ |
| $=E[flip(A+C) \mid A=1]$ | $=E[flip(A+C) \mid A=0]$ |
| $=E[flip(A)+flip(C) \mid A=1]$ | $=E[flip(C) \mid A=0]$ |
| $=E[flip(1)]+E[flip(C)]$ | $=E[flip(C)]$ |
| $=0.5+0.5*E[C]$ | $=0.5*E[C]$ |

# Counterfactual random variables

| ID | C | A | $Y^{A=1}$ | $Y^{A=0}$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | *0* |
| 2 | 0 | 1 | 1 | *0* |
| 3 | 0 | 1 | 0 | *0* |
| 4 | 0 | 0 | *1* | 1 |
| 5 | 1 | 0 | *0* | 0 |

# Counterfactual random variables

- What is $E[Y \mid A=1]$?
- What is $E[Y^{A=1} = 1]$?
- What is $E[Y^{A=1} = 1 \mid A=1]$?
- What is $E[A^{Y=1}]$?



# Counterfactual random variables

- Risk difference: $\Pr[Y^{A=1}=1]-\Pr[Y^{A=0}=1]$

- Risk ratio: $\Pr[Y^{A=1}=1] \div \Pr[Y^{A=0}=1]$

- Odds ratio: $(\Pr[Y^{A=1}=1] \div \Pr[Y^{A=1}=0])$
  $\div (\Pr[Y^{A=0}=1] \div \Pr[Y^{A=0}=0])$