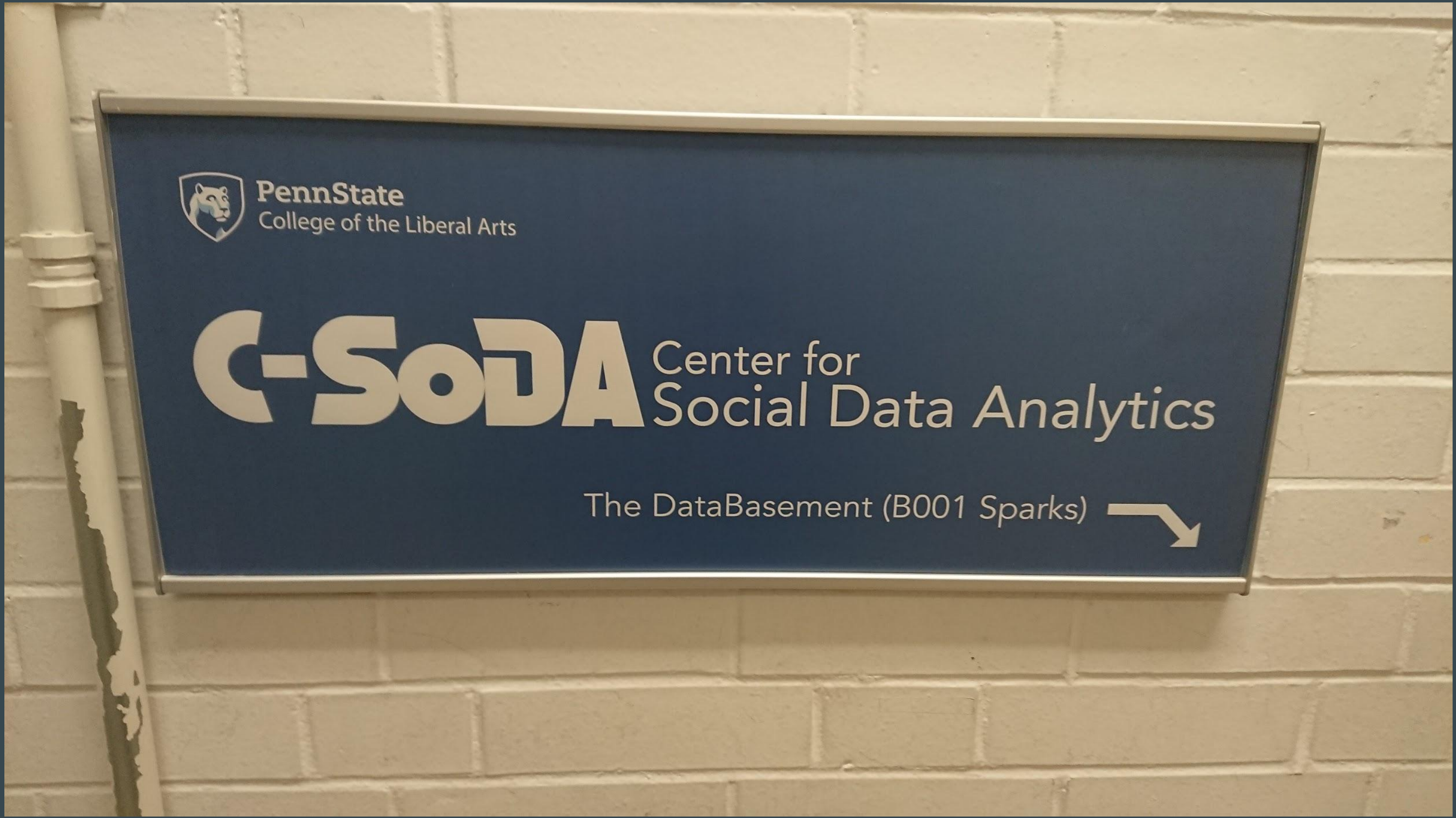
Random Variables and PMFs

28 January 2019
PLSC 309

New office sign



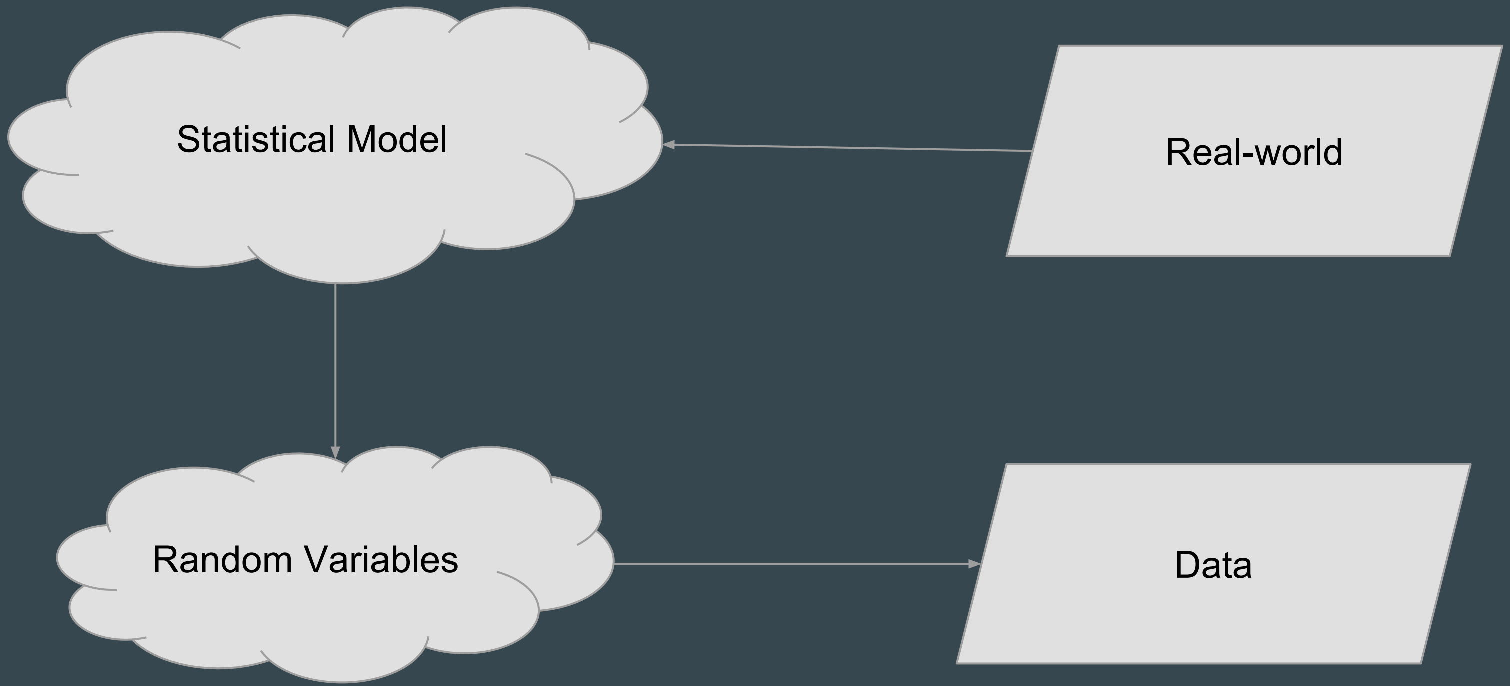
Review

- *Variables* are numeric representations of features of the real-world we're interested
- There are different numeric structures used to represent variables
  - Continuous
  - Discrete
  - Categorical
- Data is composed of *observations* of *variables*
- We're thinking probabilistically, so we realize that our data could change if we had different observations
- ...but there will be a pattern between those observations

Random variables: examples

- Age
- Gender
- Income
- State of Residence
- Political Ideology
- Temperature
- Church attendance
- Voter turnout
- Presidential vote preference
- Height and weight

Random variables are *theoretical*



Random variables take on different values

Random variables take on different values

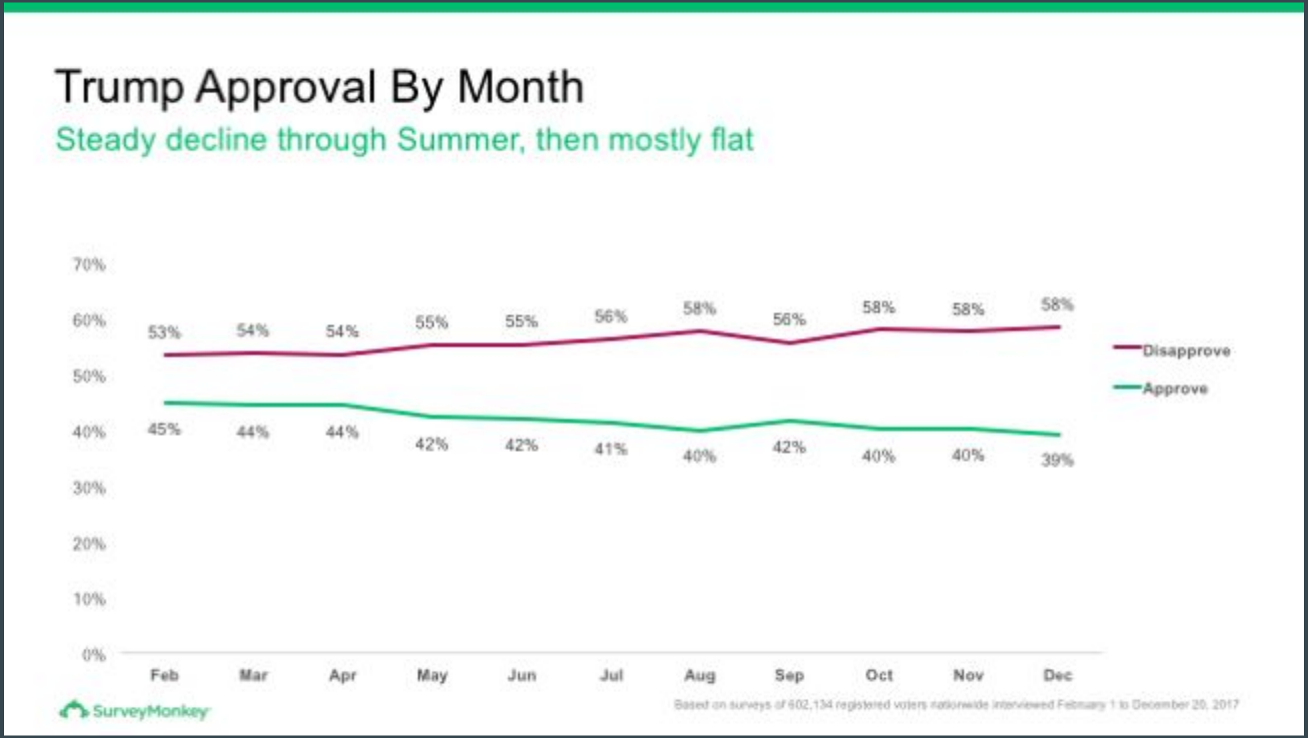
Random Variables take on different values

Day	Number of Giraffes
July 20	15
July 21	13
July 22	17
July 23	16
July 24	12
July 25	15
July 26	18

Random Variables take on different values

Day	Number of Giraffes
Dec 20	14
Dec 21	11
Dec 22	12
Dec 23	16
Dec 24	13
Dec 25	17
Dec 26	12

Random Variables take on different values



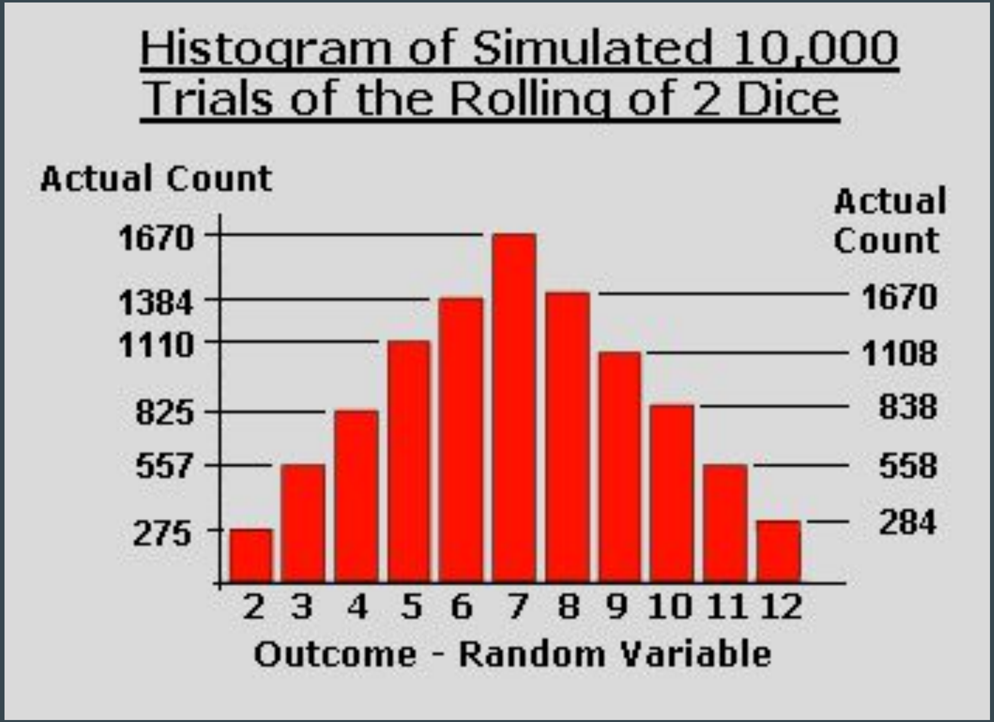
Predictability vs. Certainty

- Random variables are *predictable*
  - They follow a pattern
  - They are not *perfectly predictable*
- Random variables are *uncertain*
  - You can’t predict each individual observation
- Aggregate predictions *are* possible
  - I.e. general trends
- Specific predictions *are not* possible
  - Well, they’re possible, but they will probably be wrong

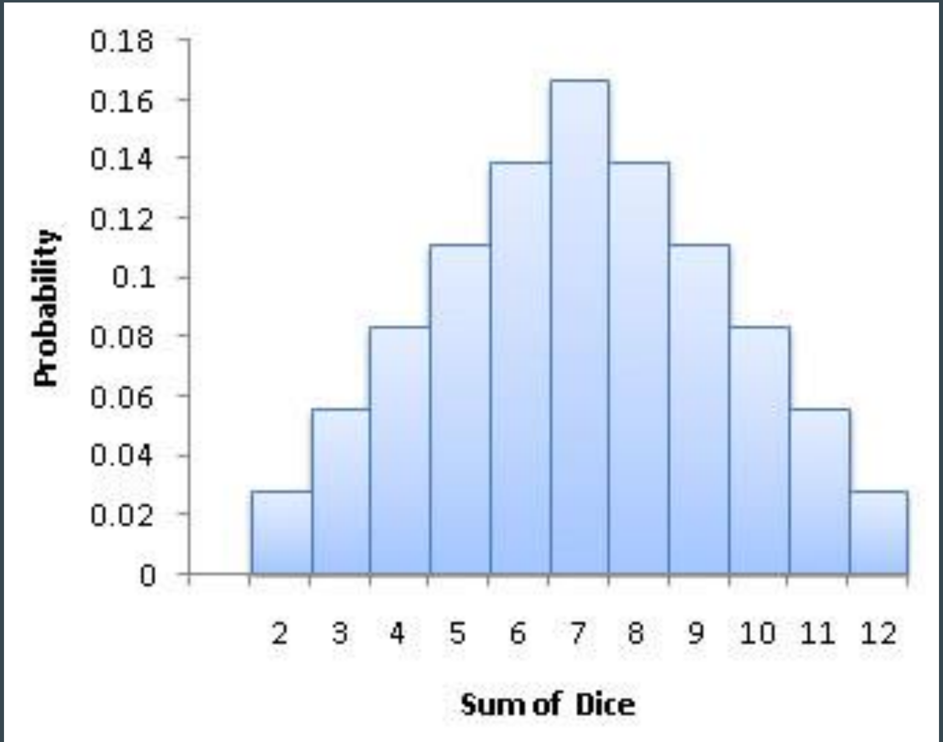
Probability Mass Function

- Represented by a special type of histogram where
  - Each “bin” is a discrete outcome
  - Y-axis is a proportion, not a count
- Represents an “infinitely” long process

Histogram - two dice



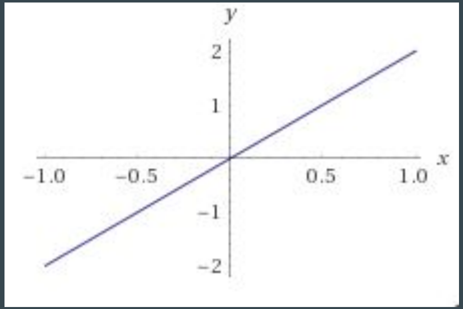
PMF - two dice


0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Probability
2
3
4
5
6
7
8
9
10
11
12
Sum of Dice

What is a function?

- A function has three parts:
  - Input
  - Transformation
  - Output
- Example: $f(x) = 2x$

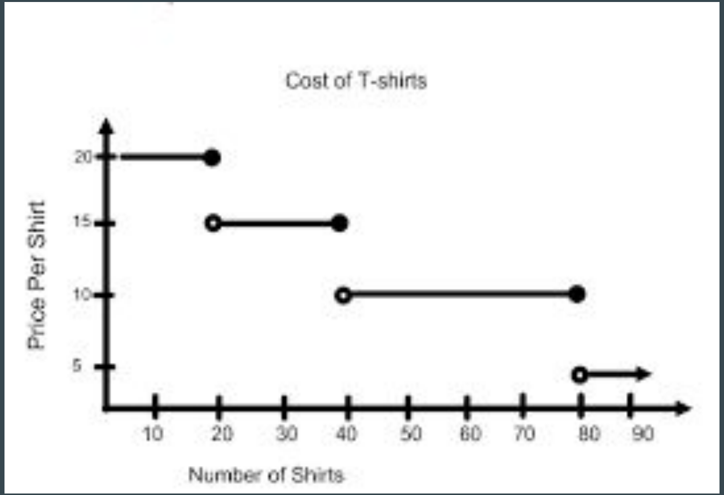
Input	Transformatio n	Output
2	2(2)	4
3	2(3)	6
4	2(4)	8



PMF is a stepwise function

- PMFs are only used for *discrete* or *ordinal* data
- Discrete or ordinal data have gaps
- Functions with gaps are known as stepwise functions

$$f(x) = \begin{cases} 20 & 0 < x < 20 \\ 15 & 20 < x < 40 \\ 10 & 40 < x < 80 \\ 5 & x > 80 \end{cases}$$



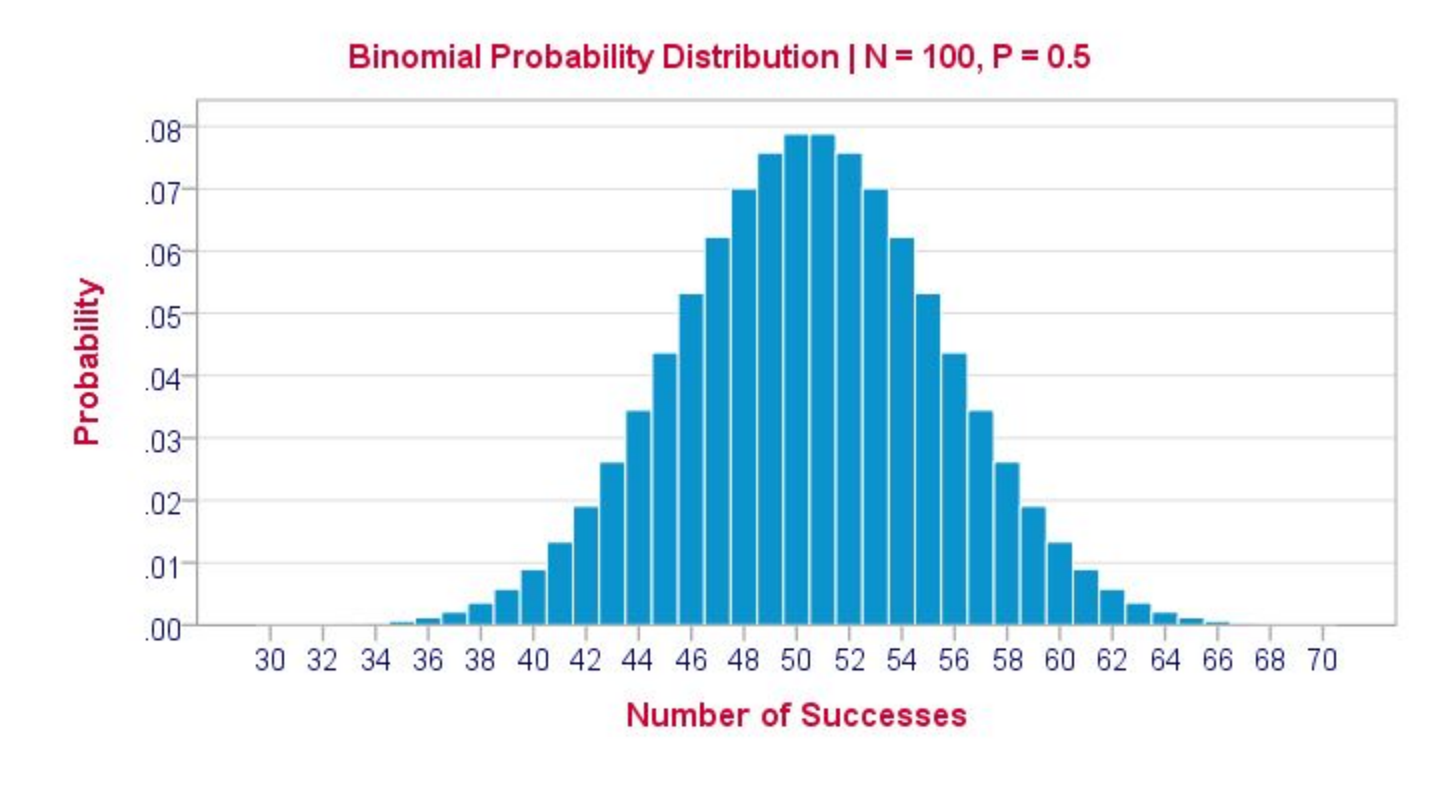
Representing PMFs

- We can represent PMFs in the following way:
    - In functional notation
    - Histogram
    - Table

Flip a coin 100 times, how many heads do you get?

Trial	Number of Heads
1	52
2	47
3	45
4	60
5	49
6	54
7	38

Flip a coin 100 times, how many heads do you get?



Flip a coin 100 times, how many heads do you get?

Number of heads	Probability
1	.0000001
2	.00000013
3	.00000014
…	…
50	.08
…	…
99	.0000001

Expected value = Mean of PMF

- Called expected, because it’s based on infinite trials
- What’s the most likely value of a variable?
- What’s the average value of a million repeated trials?
- In statistics, we assume the above to questions are really the same

Formula for expected value

Expected value of a Discrete Random Variable

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) &= x_1 \times P(X = x_1) + \cdots + x_k \times P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \qquad (2.71) \end{aligned}$$

The Greek letter μ may be used in place of the notation $E(X)$.

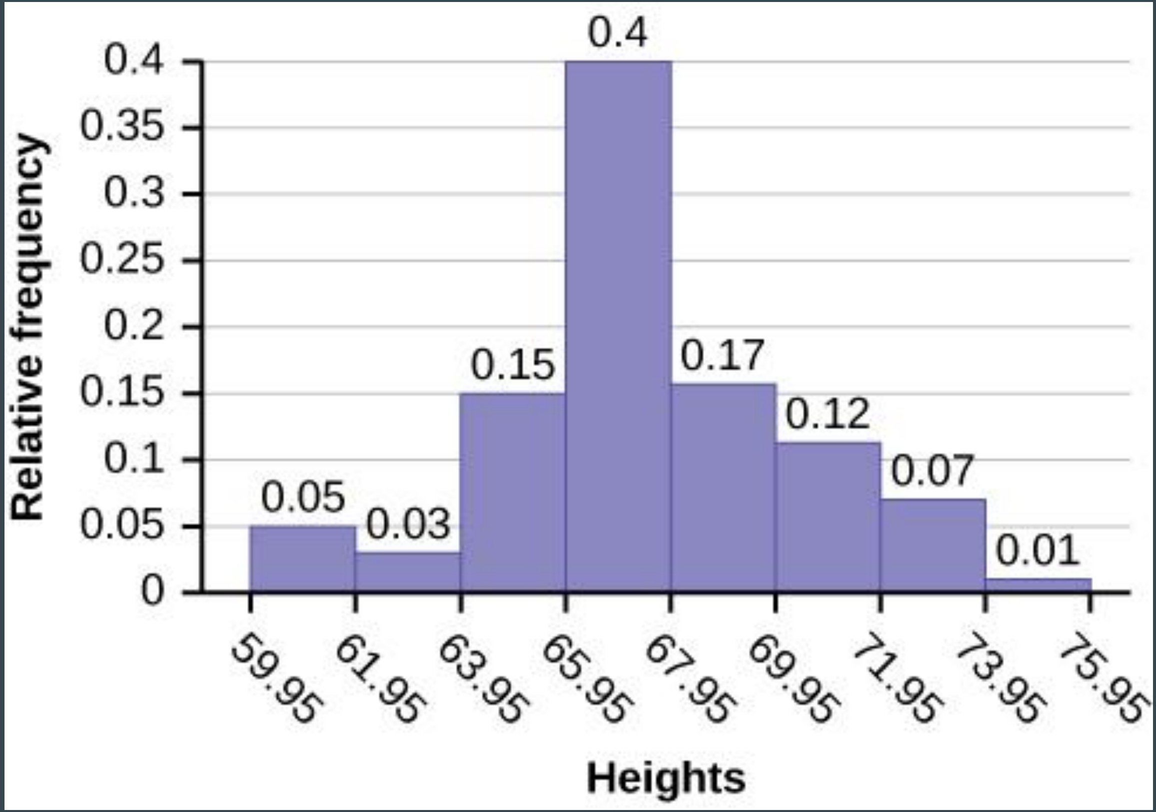
Calculating EV for PMF

Number of seats R's gain in Senate	Probability
0	0.10
1	0.30
2	0.60

Calculating EV for PMF

- $P(X=x_i)$
  - $P(X=0)$
  - $P(X=1)$
  - $P(X=2)$
- $X_1 * P(X=1) + X_2 * P(X=2) + X_3 * P(X=3)$
- $0 * .1 + 1 * .3 + 2 * .6 = 1.5$

Calculating EV of PMF



Calculating EV for PMF

Height	P(Height)
59.95	0.05
61.95	0.03
63.95	0.15
65.95	0.4
67.95	0.17
69.95	0.12
71.95	0.07
73.95	0.01

Calculating EV for PMF

- $P(X=X_i)$
  - $P(X=59.95) = 0.05$
  - ...
  - $P(X=73.95) = 0.01$
- $59.95(0.05) + 61.95(0.03) + ... + 73.95(0.01)$

Variance of a PMF

- We have learned to calculate E(X), which is also the mean of X
- Earlier, we defined the distance as the *squared difference* between a single value (X_i) and the mean (μ)
- And the variance is defined as the average of the distances

Variance of a PMF

General variance formula

If X takes outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by $Var(X)$ or the symbol σ^2, is

$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 \times P(X = x_1) + \cdots \\ &\qquad \cdots + (x_k - \mu)^2 \times P(X = x_k) \\ &= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j) \qquad (2.72)\end{aligned}$$

The standard deviation of X, labeled σ, is the square root of the variance.

Calculating variance of PMF

Number of seats R’s gain in Senate	Probability
0	0.10
1	0.30
2	0.60

Calculating Var(X)

- $\mu = 1.5$
- $(x_i - \mu)^2$
  - $(X_1 - 1.5)^2 = 2.25$
  - $(X_2 - 1.5)^2 = 0.25$
  - $(X_3 - 1.5)^2 = 2.25$
- $(x_i - \mu)^2 * P(x_i)$
  - $2.25 * .1 = .225$
  - $0.25 * .3 = .075$
  - $2.25 * .6 = 1.35$
- $Var(X) = 2.325$

Review

- Random variables are a theoretical concept
  - They are quantified phenomena
  - Repeated infinitely (theoretically)
- Discrete or ordinal variables follow a probability mass function
  - Each possible value represented by its own probability
  - All probabilities sum to 1
- The expected value of a PMF is its mean
- Calculation for variance of a random variable