SOC 3811/5811:
BASIC SOCIAL STATISTICS

Discrete Random Variables

Probability

Random Event

An event in which the outcome is determined by chance

Probability

An expression of how likely it is that a particular event (E) will result from a random event

Probabilities are expressed as a number between 0 (the event will certainly not occur) and 1 (the event certainly will occur)

P(E)=0.2 means “the probability that event X will occur is 0.2”

P(~E)=0.8 means “the probability that event X will not occur is 0.8”

P(E) + P(~E) = 1.0 … so … P(~E) = 1.0 – P(E)

Probability

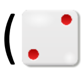
Random Event: *Tossing a Fair Coin*

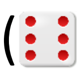
Possible Outcomes:  

Probabilities: P() = 0.5 ; P() = 0.5 ; P() + P() = 1.0

Random Event: *Roll a Pair of Dice*

Possible Outcomes:     

Probabilities: P() = 0.167 ; P() = 0.167; … ; P() = 0.167

P() + P() + … + P() = 1.0

Random Event: *Whether a Person Has Ever Been Arrested*

Possible Outcomes: Has Been Arrested; Has Never Been Arrested

Probabilities: P(Ever Arrested) = 0.3 ; P(Never Arrested) = 0.7

P(Ever Arrested) + P(Never Arrested) = 1.0

Probability

Independent Random Events

Events A and B are *independent* if P(A) is the same regardless of whether Event B happens, and P(B) is the same regardless or whether Event A happens

Example: If you flip two coins, P() = 0.5 for the second coin regardless of what happened when you flipped the first coin

Conditional (or non-Independent) Random Events

Event B is conditional on Event A if P(B) depends on whether or not Event A happened

Example: If you draw two cards from a deck of playing cards, P() on your second card depends on what you drew on your first card.

A person’s birthday is a random event. If you pick two people at random, are their birthdays **independent** or **dependent** events?

WORKSHEET

Whether a job applicant gets the job is a random event. If you pick two people from among the applicants for a particular job, and only one of them can get the job, are their application outcomes **independent** or **dependent** events?

Probability

Independent Random Events

For two independent random events A and B…

…the probability that both Event A and Event B happen is:

P(A and B) = P(A) x P(B)

Example: P(and) = P() x P() = 0.167 x 0.167 = 0.028

…the probability that either Event A or Event B happens is:

P(A or B) = P(A) + P(B) – P(A and B)

Example: P( or) = P() + P() – P( and )

0.167 + 0.167 – 0.028 = 0.361

Probability

Independent Random Events

For two independent random events A and B…

…the probability that both Event A and Event B happen is:

P(A and B) = P(A) x P(B)

Example: P(and ) = P() x P() = 0.167 x 0.500 = 0.083

…the probability that either Event A or Event B happens is:

P(A or B) = P(A) + P(B) – P(A and B)

Example: P( or ) = P() + P() – P( and )

0.167 + 0.500 – 0.083 = 0.584

Probability

Conditional (or non-Independent) Random Events

If P(B) for Event B is dependent upon the outcome of Event A, then…

…the probability that both Event A <u>and</u> Event B happen is:

$P(A \text{ and } B) = P(A) \times P(B|A)$

Example: What is the probability of drawing two cards (A and B) from a deck and having them both be red? For the first drawn card, P(A) = 26/52 = 0.5. However, for the second drawn card, P(B) is not 26/52 because there are only 51 cards left in the deck. The probability that the second card is red *given that the first card was red* … P(B|A) … is 25/51 = 0.4902. So P(A and B) = P(A) x P(B|A) = 0.5 x 0.4902 = 0.2451

< < < Dealers Up Card > > >

Your Hand	2	3	4	5	6	7	8	9	10	A
8 or less	H	H	H	H	H	H	H	H	H	H
9	H	D	D	D	D	H	H	H	H	H
10	D	D	D	D	D	D	D	D	H	H
11	D	D	D	D	D	D	D	D	D	H
12	H	H	S	S	S	H	H	H	H	H
13	S	S	S	S	S	H	H	H	H	H
14	S	S	S	S	S	H	H	H	H	H
15	S	S	S	S	S	H	H	H	H/SU	H
16	S	S	S	S	S	H	H	H/SU	H/SU	H/SU
17+	S	S	S	S	S	S	S	S	S	S
A8-10	S	S	S	S	S	S	S	S	S	S
A7	S	D	D	D	D	S	S	H	H	H
A6	H	D	D	D	D	H	H	H	H	H
A5	H	H	D	D	D	H	H	H	H	H
A4	H	H	D	D	D	H	H	H	H	H
A3	H	H	H	D	D	H	H	H	H	H
A2	H	H	H	D	D	H	H	H	H	H
A-A,8-8	SP	SP	SP	SP	SP	SP	SP	SP	SP	SP
10-10	S	S	S	S	S	S	S	S	S	S
9-9	SP	SP	SP	SP	SP	S	SP	SP	S	S
7-7	SP	SP	SP	SP	SP	SP	H	H	H	H
6-6	H	SP	SP	SP	SP	H	H	H	H	H
5-5	D	D	D	D	D	D	D	D	H	H
4-4	H	H	H	H	H	H	H	H	H	H
3-3,2-2	H	H	SP	SP	SP	SP	H	H	H	H
	2	3	4	5	6	7	8	9	10	A

< < < Dealers Up Card > > >

H = Hit S = Stand D = Double Down SP = SPlit

H/SU = SUrrender if able to, otherwise Hit

Probability Distributions

If we flip a fair coin 3 times, we will observe independent events A, B, and C. What are the possible outcomes? How likely is each outcome?

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

P() = 0.5 x 0.5 x 0.5 = 0.125

Probability Distributions

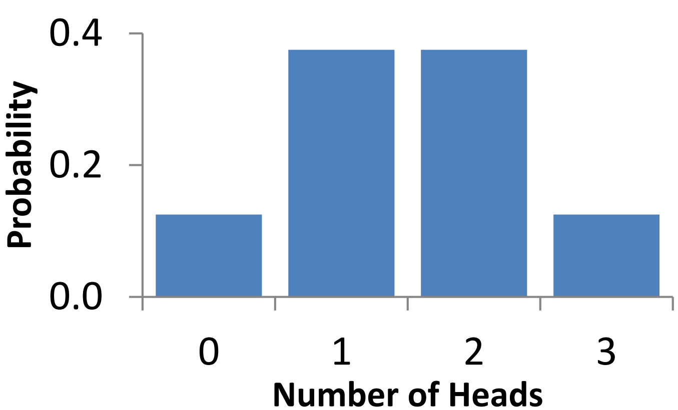
If we flip a fair coin 3 times, how many different numbers of "heads" might come up? How likely is each possible value?

P(3 Heads) = P() = 0.125

P(2 Heads) = P() + P() + P() = 3x0.125 = 0.375

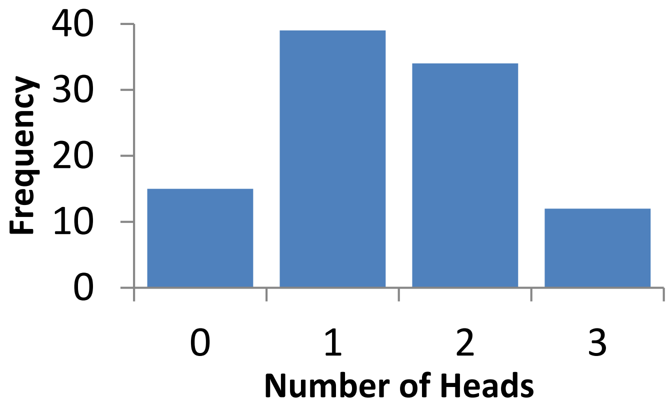
P(1 Heads) = P() + P() + P() = 3x0.125 = 0.375

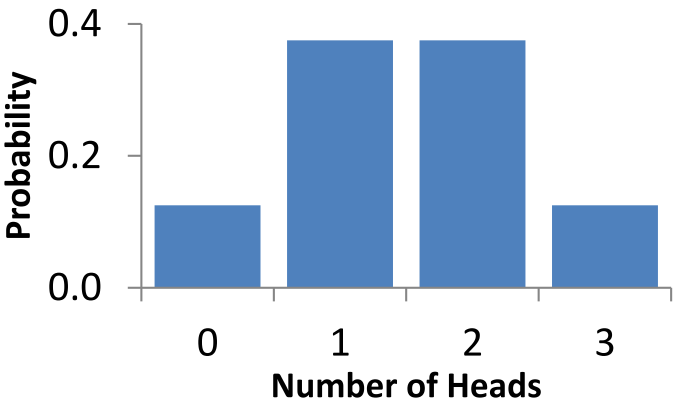
P(0 Heads) = P() = 0.125



Probability Distributions

Now imagine I went and *actually* flipped three fair coins, recorded the number of heads, and repeated that game 100 times. On the left is what I observed:





What we **saw**

This is an **observed** distribution

(from flipping 3 coins 100 times)

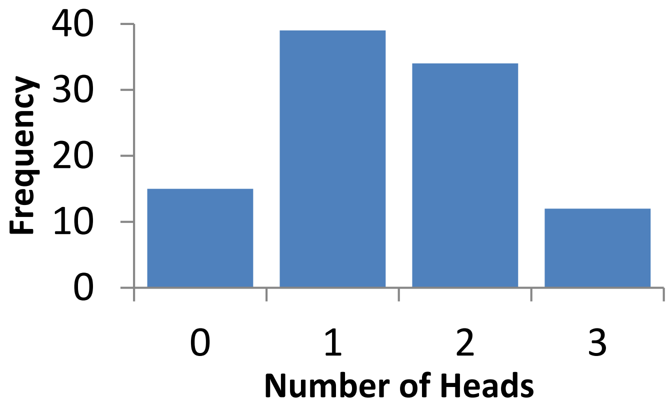
What we **expected**

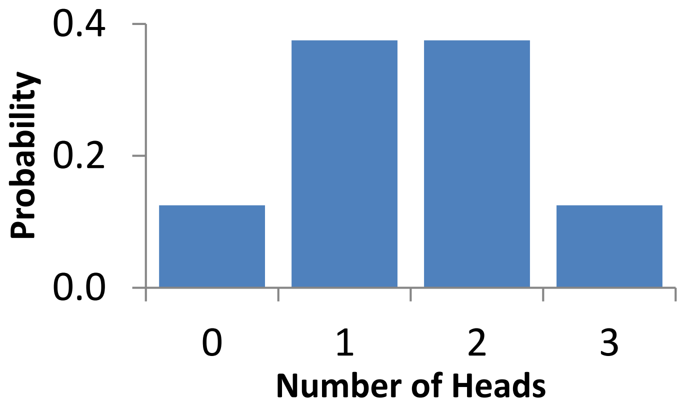
This is a **theoretical** distribution

(if we flipped 3 coins ∞ times)

Probability Distributions

Now imagine I went and *actually* flipped three fair coins, recorded the number of heads, and repeated that game 100 times. On the left is what I observed:





What we **saw**

This is **sample** data
(from flipping 3 coins 100 times)

What we **expected**

This is **population** data
(if we flipped 3 coins ∞ times)

Random Variables

Random variable

The numeric outcome of a random event

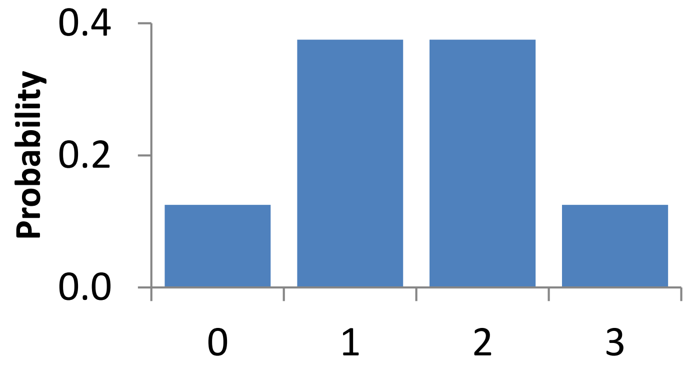
Random Event	**Random Variable**
Flip a fair coin 3 times	Number of "heads" observed
Sample a person from the U.S.	Whether they have been arrested
Sample water from a lake	Bacteria per cubic inch

For each random variable, the outcome values vary across cases as a function of a chance process, and are unknown in advance

Random Variables

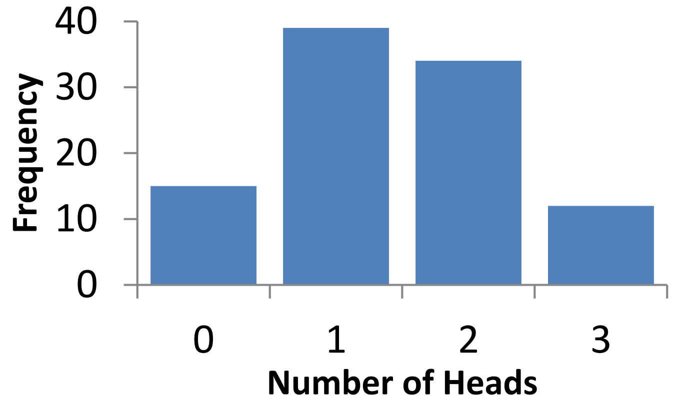
Random Event
3 Coin Flips
(Repeated Infinitely)

Theoretical Distribution
for the **Population** of
All Possible 3 Coin Flips



Random Event
3 Coin Flips
(Repeated 100 Times)

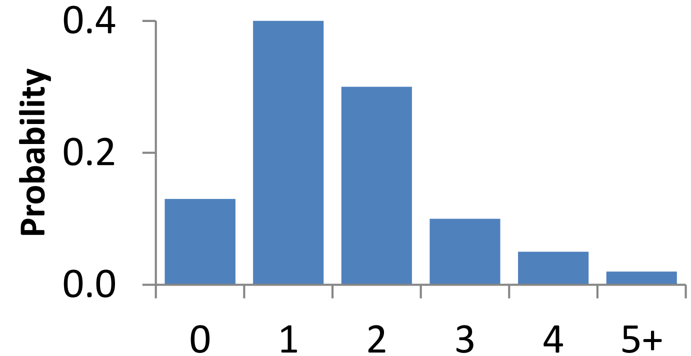
Observed Distribution
for a **Sample** of
100 Sets of 3 Coin Flips



Random Variables

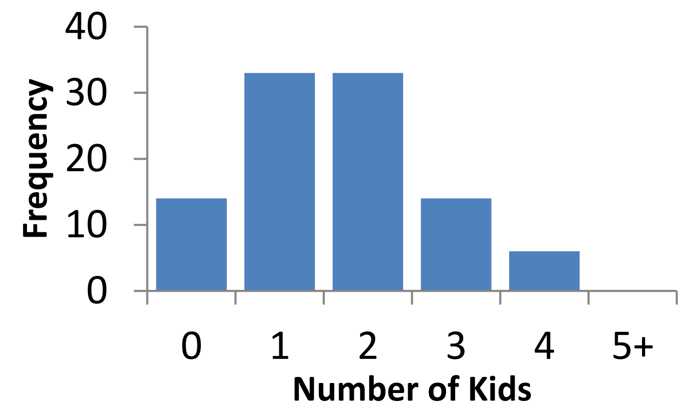
Random Event
Survey Question About # of Kids
(Repeated Infinitely)

Theoretical Distribution
for the **Population** of
All Possible Responses



Random Event
Survey Question About # of Kids
(Repeated 100 Times)

Observed Distribution
for a **Sample** of
100 Sets of Responses

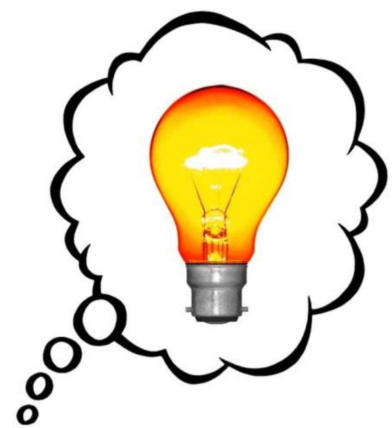


Reality

Data

2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486

Knowledge

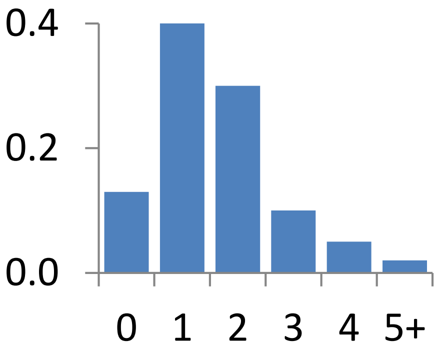


Reality

Data

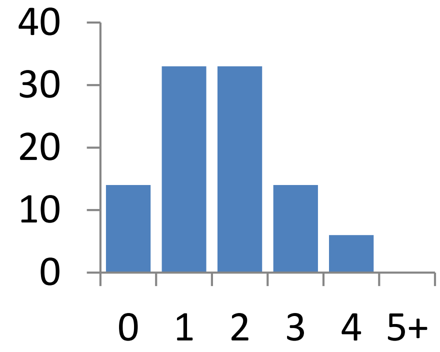
Knowledge

Theoretical Distribution
for the **Population** of
All Possible Responses



2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486

Observed Distribution
for a **Sample** of
100 Sets of Responses



Random Variables

Today

We <u>know</u> the actual probabilities associated with the random event that generates the theoretical distribution (e.g., coin flips)

We will learn to *describe* and *make practical use* of these theoretical distributions

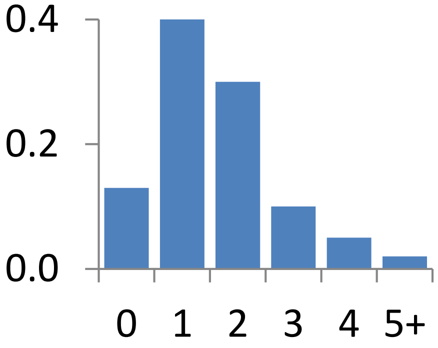
Later (and in Life)

We <u>do not know</u> the actual probabilities associated with the random event (e.g., number of children per person, which candidate will win)

Our ability to *describe* and *make practical use* of theoretical distributions will allow us to <u>infer</u> those probabilities

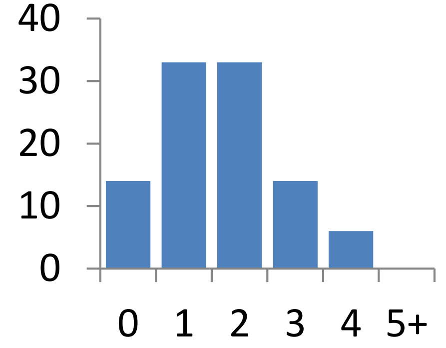
Reality | Data | Knowledge

Theoretical Distribution for the **Population** of **All** Possible Responses



2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486

Observed Distribution for a **Sample** of **100** Sets of Responses

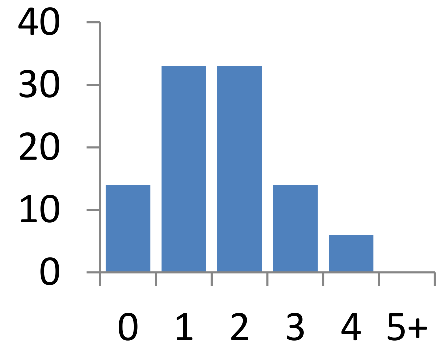


Reality | Data | Knowledge

?

2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486

Observed Distribution
for a **Sample** of
100 Sets of Responses



Random Variables

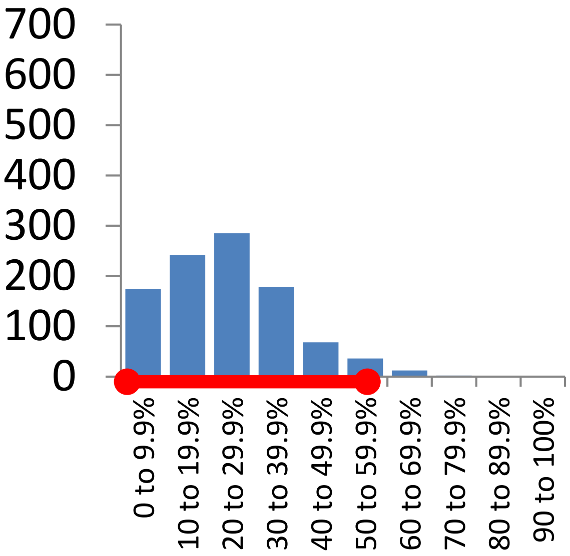
1. Confidence Intervals

Based on the distribution of Y in sample data, we are confident that the distribution of Y in the population has particular qualities (e.g., that its mean is within a certain range of values)

"With 95% certainty, I conclude based on my sample data that between 25% and 35% of everyone in the population has been arrested"

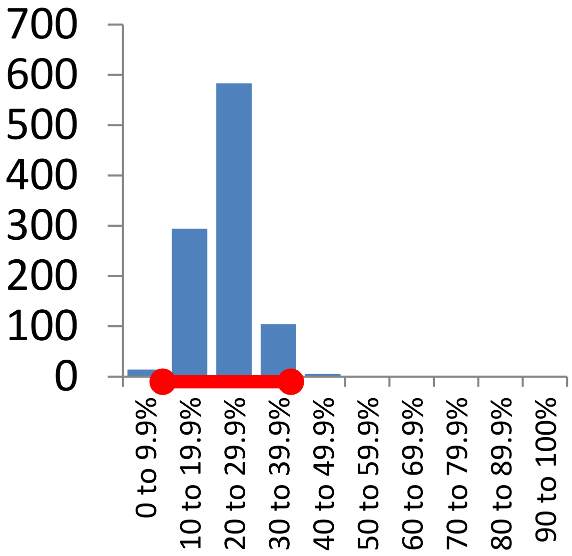
Random Variables

<u>Question</u>: What percentage of Americans were high school graduates in 1940?



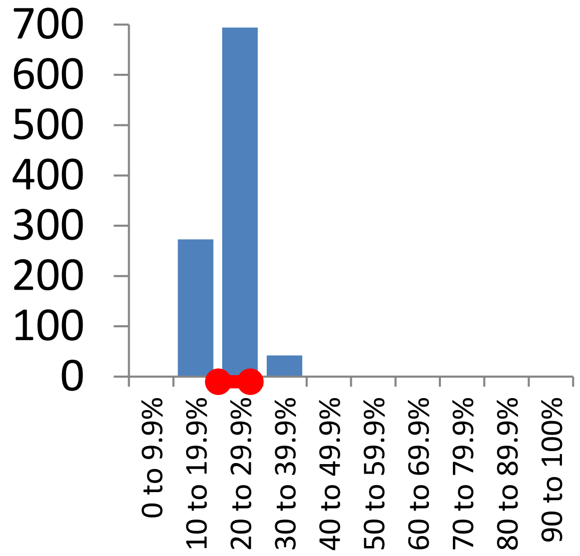
Results from asking **10** randomly selected people; trial repeated 1,000 times

$$\pm\frac{1}{\sqrt{10}}\times 100\% = \pm 32\%$$



Results from asking **50** randomly selected people; trial repeated 1,000 times

$$\pm\frac{1}{\sqrt{50}}\times 100\% = \pm 14\%$$



Results from asking **1,000** randomly selected people; trial repeated 1,000 times

$$\pm\frac{1}{\sqrt{1000}}\times 100\% = \pm 3\%$$

Random Variables

2. Hypothesis Tests

Based on the distribution of Y in the sample data, we can evaluate the likely truth of theoretically-informed hypotheses about the distribution of Y in the population (e.g., that the mean of X is above some value)

"With 95% certainty, I reject the claim that fewer than 20% of everyone in the population has ever been arrested"

Random Variables

Imagine you flip a bottle cap 100 times, and count the number of heads

If you got 52 heads out of 100, would you reject the idea that P(Heads) = 0.5?

What if you got 75 heads?

Random Variables

Discrete Random Variable

Can only equal a finite number of distinct values

Example: When you flip a coin 3 times, you can only get four possible values … the whole numbers 0 through 3

Continuous Random Variable

Can take any numeric value within a range of values

Example: The number of miles from campus students live can take on just about any value (0.12, 1.17, 2.00, etc.)

Note: Discrete random variables with lots of values (e.g., number of Facebook friends) are often treated as continuous; continuous variables that are rounded (e.g., age) may seem discrete

Discrete Random Variables

Random variable Y has k possible values

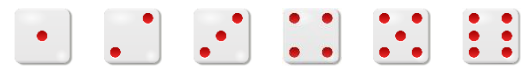
P(Y=1) + P(Y=2) + … P(Y=k) = 1.0

Example: If I roll a single die (the random event) and record the number of pips that come up (Y), then there are 6 possible value that might come up: ⚀ ⚁ ⚂ ⚃ ⚄ ⚅

P(Y=⚀) + P(Y=⚁) + … + P(Y=⚅) = 1.0

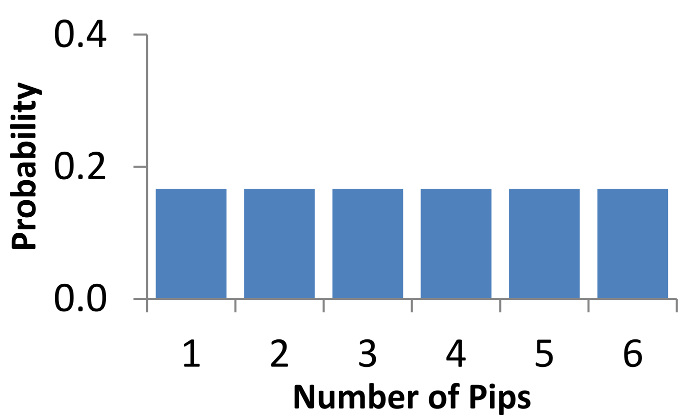
0.167 + 0.167 + 0.167 + 0.167 + 0.167 + 0.167 = 1.0

Discrete Random Variables

Example: If I roll a single die (the random event) and record the number of pips that come up (Y), then there are 6 possible value that might come up: 

The **probability distribution function** for Y is:

k	1	2 	3 	4 	5 	6 
p(Y=k)	0.167	0.167	0.167	0.167	0.167	0.167

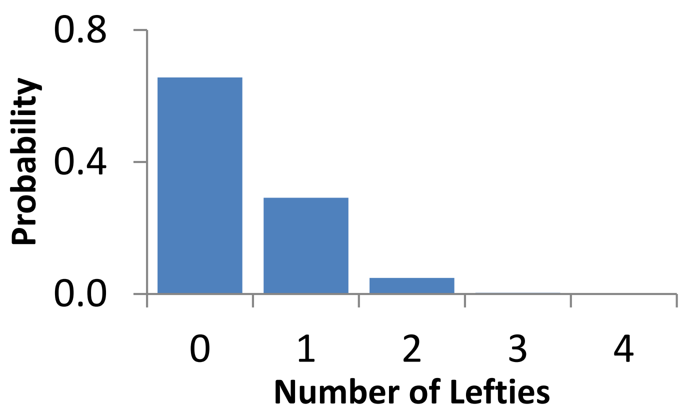


Discrete Random Variables

Example: The probability that a baby is born left-handed is 0.1. For families with 4 babies, the number of left-handed babies (Y) can thus be 0, 1, 2, 3, or 4

The **probability distribution function** for Y is:

k	0	1	2	3	4
p(Y=k)	0.6561	0.2916	0.0486	0.0036	0.0001



Discrete Random Variables

Where is the center of the theoretical distribution of discrete random variable Y?

The **expected value** of Y … called E(Y) or μ_Y … is the mean of Y in the population of all possible outcomes

One way to think about it: If you repeated the random event an infinite number of times, the expected value is the mean of Y that you would observe

Discrete Random Variables

For discrete random variable Y:

$Y_1, Y_2, Y_3, \ldots, Y_k$ are all of the k possible values of Y

$p_1, p_2, p_3, \ldots, p_k$ are the probabilities of observing each value

The **expected value** of Y … the mean of Y that you would get if you repeated the random trial an infinite number of times … is:

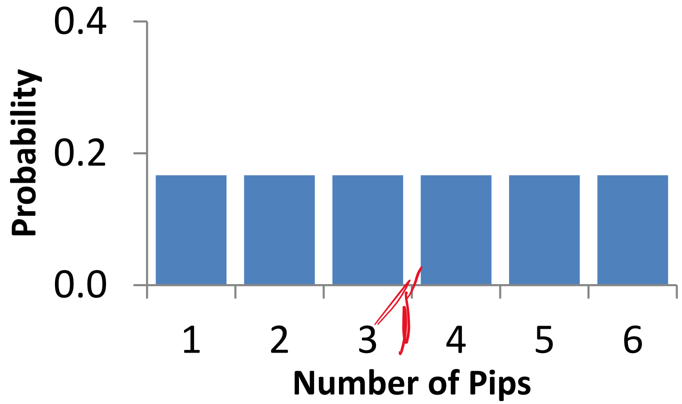
$$E(Y) = \mu_Y = \sum_{i=1}^{k} Y_i p(Y_i)$$

Discrete Random Variables

Example: If I roll a single die (the random event) and record the number of pips that come up (Y), then there are 6 possible value that might come up:

The **probability distribution function** for Y is:

k	1	2	3	4	5	6
p(Y=k)	0.167	0.167	0.167	0.167	0.167	0.167



$$E(Y) = \mu_Y = \sum_{i=1}^{k} Y_i p(Y_i)$$

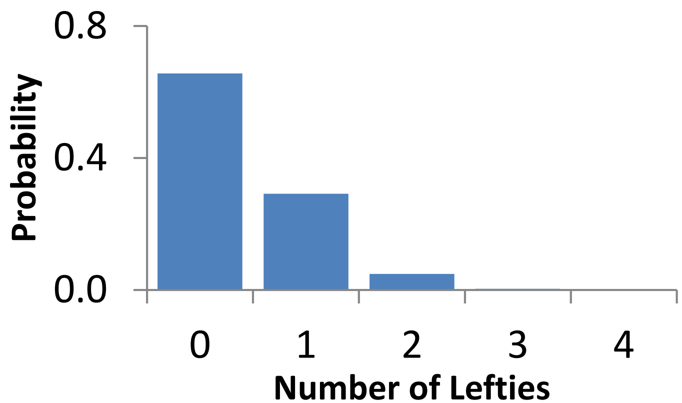
$$E(Y) = (1)(0.167) + (2)(0.167) + (3)(0.167) + (4)(0.167) + (5)(0.167) + (6)(0.167) = 3.5$$

Discrete Random Variables

Example: The probability that a baby is born left-handed is 0.1. For families with 4 babies, the number of left-handed babies (Y) can thus be 0, 1, 2, 3, or 4

The **probability distribution function** for Y is:

k	0	1	2	3	4
p(Y=k)	0.6561	0.2916	0.0486	0.0036	0.0001



$$E(Y) = \mu_Y = \sum_{i=1}^{k} Y_i p(Y_i)$$

$$E(Y) = (0)(0.6561) + (1)(0.2916) + (2)(0.0486) + (3)(0.0036) + (4)(0.0001) = 0.4$$

Discrete Random Variables

E(Y) is a measure of central tendency for the distribution of theoretical discrete variable Y

How much "spread" or variability is there?

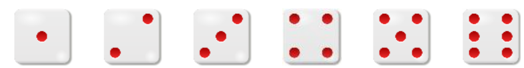
For discrete random variable Y, the variance of Y … σ^2_Y … is

$$\sigma_Y^2 = \sum_{i=1}^{k} \left(Y_i - \mu_Y\right)^2 p(Y_i)$$

and the standard deviation of Y … σ_Y … is

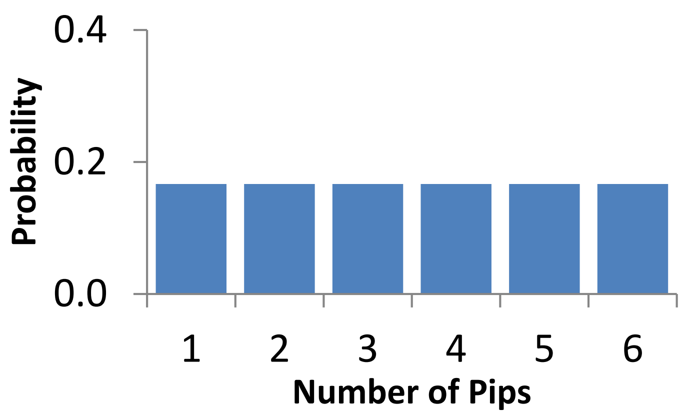
$$\sigma_Y = \sqrt{\sum_{i=1}^{k} \left(Y_i - \mu_Y\right)^2 p(Y_i)}$$

Discrete Random Variables

Example: If I roll a single die (the random event) and record the number of pips that come up (Y), then there are 6 possible value that might come up: 

The **probability distribution function** for Y is:

k	1	2 	3 	4 	5 	6 
p(Y=k)	0.167	0.167	0.167	0.167	0.167	0.167



$$\sigma_Y^2 = \sum_{i=1}^{k} \left(Y_i - \mu_Y\right)^2 p(Y_i)$$

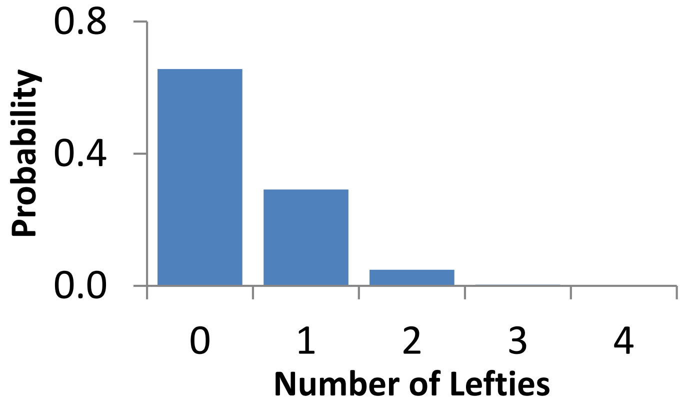
$$\sigma^2_Y = (1-3.5)^2(0.167) + (2-3.5)^2(0.167) + (3-3.5)^2(0.167) + (4-3.5)^2(0.167) + (5-3.5)^2(0.167) + (6-3.5)^2(0.167) = 2.92$$

Discrete Random Variables

Example: The probability that a baby is born left-handed is 0.1. For families with 4 babies, the number of left-handed babies (Y) can thus be 0, 1, 2, 3, or 4

The **probability distribution function** for Y is:

k	0	1	2	3	4
p(Y=k)	0.6561	0.2916	0.0486	0.0036	0.0001



$$\sigma_Y^2 = \sum_{i=1}^{k} \left(Y_i - \mu_Y\right)^2 p(Y_i)$$

$\sigma^2_Y = (0-0.4)^2(0.6561) + (1-0.4)^2(0.2916) + (2-0.4)^2(0.0486) + (3-0.4)^2(0.0036) + (4-0.4)^2(0.0001) = 0.36$

Discrete Random Variables

For the discrete random variable Y = "number of babies out of 4 born left handed" we found that the expected number of left handed babies was $E(Y)=\mu_Y= 0.4$, with standard deviation $\sigma^2_Y=0.36$

If we sampled an infinite number of families with four kids and observed Y for each, the mean of Y would be 0.4

We wouldn't always (or ever) see exactly 0.4 left-handed kids; σ^2_Y indicates how much the number of left-handed kids would vary from family to family

Worksheet

Below is a probability distribution function for discrete variable Y. Compute the mean and standard deviation of this distribution

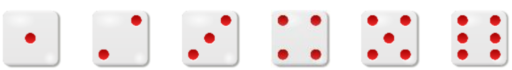
k	0	1	2	3
$P(Y=k)$	0.8	0.1	0.05	0.05

Binomial Random Variables

A special kind of discrete random variable … called a binomial random variable … is one in which the random event generating it has only two possible outcomes

Discrete Random Variable

Random event has a finite number of distinct values; the random variable is a count of how often each value occurs.

Example: 

Binomial Random Variable

Random event has exactly two possible values (“success” and “failure”); the random variable is a count of how often “success” occurs over *n* repeated trials. *Example*: 

Binomial Random Variables

Binomial random variable Y equals the number of "successes" out of n binomial events when the probability of "success" in any one event is p

The probability that Y = k (where k ranges from 0 to n) is:

$$P(Y=k)=\frac{n!}{k!(n-k)!}p^{k}(1-p)^{n-k}$$

For example, what is the probability of getting k=2 heads out of n=3 flips of a fair coin where P(Heads)=0.5?

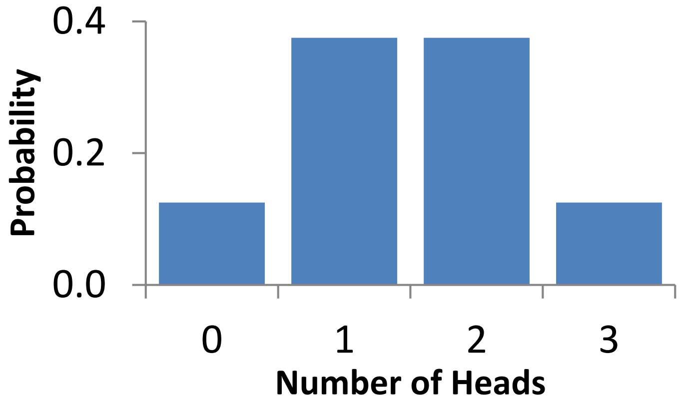
$$P(Y=2)=\frac{3!}{2!(3-2)!}0.5^{2}(1-0.5)^{3-2}=\frac{6}{2\,x\,1}(0.25)(0.5)$$

$$=0.375$$

Binomial Random Variables

Applying this formula to each possible value of Y when n=3 and p=0.5, we get the probability distribution function:

k	0	1	2	3
$P(Y=k)$	0.125	0.375	0.375	0.125



Binomial Random Variables

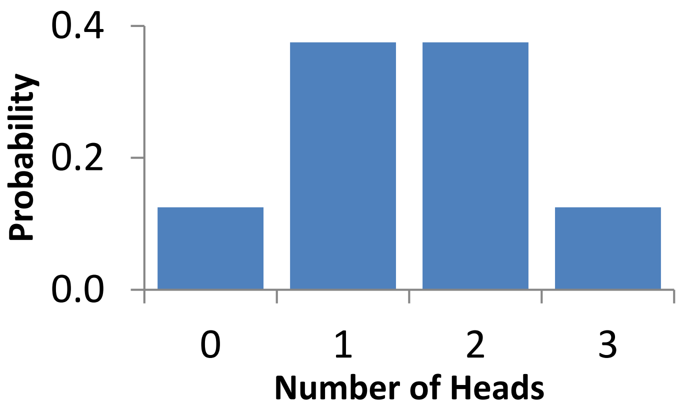
But we already knew this from before…

P(3 Heads) = P() = 0.125

P(2 Heads) = P() + P() + P() = 3x0.125 = 0.375

P(1 Heads) = P() + P() + P() = 3x0.125 = 0.375

P(0 Heads) = P() = 0.125



Binomial Random Variables

For binomial random variable Y based on n trials with probability of "success" on any given trial equal to p, the expected value of Y (or μ_Y) equals

$$E(Y) = \mu_Y = np$$

the variance σ^2_Y equals

$$\sigma_Y^2 = np(1-p)$$

and the standard deviation σ_Y equals

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{np(1-p)}$$

Binomial Random Variables

Because a binomial random variable is just a special case of discrete random variables, the formula used to compute the mean and variance of a discrete random variable…

$$E(Y) = \mu_Y = \sum_{i=1}^{k} Y_i p(Y_i)$$

$$\sigma_Y^2 = \sum_{i=1}^{k} (Y_i - \mu_Y)^2 p(Y_i)$$

…will also work for binomial random variables (if you are inclined to do extra math)

Binomial Random Variables

Whether a newborn baby is a boy is a binomial experiment with p = 0.512

For the binomial random variable "number of boys" based on n = 100 randomly selected births:

$$E(Y) = \mu_Y = np = 100 \text{ x } 0.512 = 51.2$$

$$\sigma_Y^2 = np(1-p) = 100(0.512)(1-0.512) = 24.99$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{np(1-p)} = \sqrt{24.99} = 4.999$$

If we sampled 100 births and counted the number of boys … and then repeated that experiment an infinite number of times…we would average 51.2 boys with σ_Y = 4.999

Worksheet

The probability that a new car is a “lemon” is 0.01

A factory produces 8 cars per hour

- What is the probability that the factory will produce exactly zero “lemons” in a given hour?

- What is the mean and standard deviation of the number of “lemons” the factory produces per hour?

Want More?

This is a good reading:

http://www.stat.auckland.ac.nz/~wild/ChanceEnc/Ch05.pdf

David Lane's Book

http://onlinestatbook.com/2/probability/probability.html

Gerard Dallal's Book

http://www.jerrydallal.com/LHSP/prob.htm