
INFO 2950: Intro to Data Science

Lecture 16
2023-10-23

Agenda

1. Binomial distributions
2. Explanations with distributions
 - a. Tennis example
 - b. Hypothesis testing

The Binomial Distribution

Counting the number of **positive events X** out of **total events N** where each event has **probability p** to be positive is the pattern for a binomial distribution

We'll start by defining the **probability** of a sequence, then show how to relate this to a count

Example: binary variable

- Event space: 0/1
- Parameters: p = probability of 1
- Probability function: p for 1, $(1-p)$ for 0

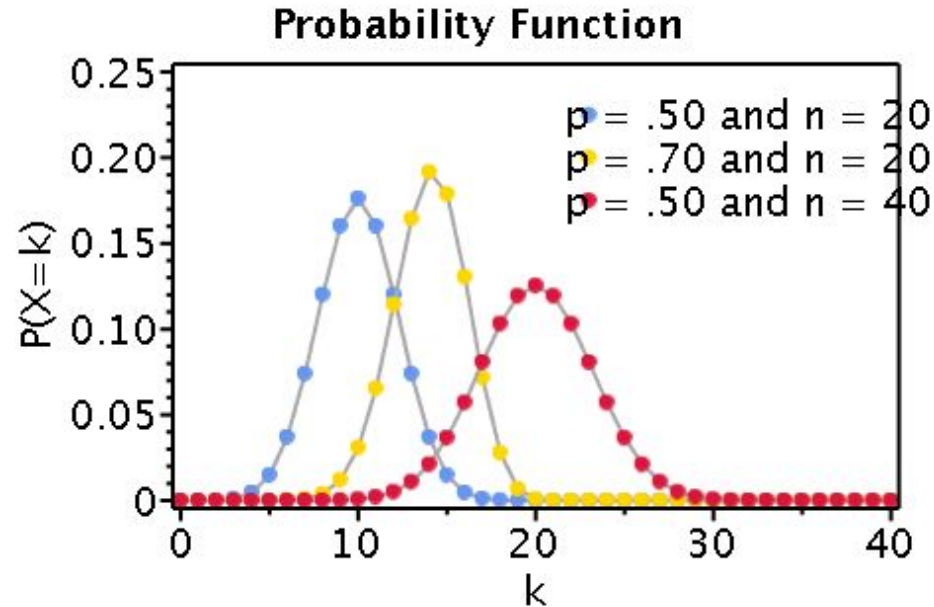


[https://en.wikipedia.org/wiki/Penny_\(United_States_coin\)](https://en.wikipedia.org/wiki/Penny_(United_States_coin))

Example: opinion poll, single yes/no question

- Event space: 0 ... N yeses
- Parameters: N, p
- Probability function: from the **Binomial distribution** $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

Counting i.i.d. events is the basis for binomial distribution



What about bootstrap?

- We can always use bootstrap samples, but there are certain processes that occur frequently where we can define properties of that process using a **probability distribution**.
- Why might we prefer using a probability distribution over bootstrapping? It's **faster** and easier than bootstrapping, and there's a lot of **theory** behind it to describe important data qualities.

Motivation: Binomial distribution statistics

$$\mathbb{E}[X] = N p$$

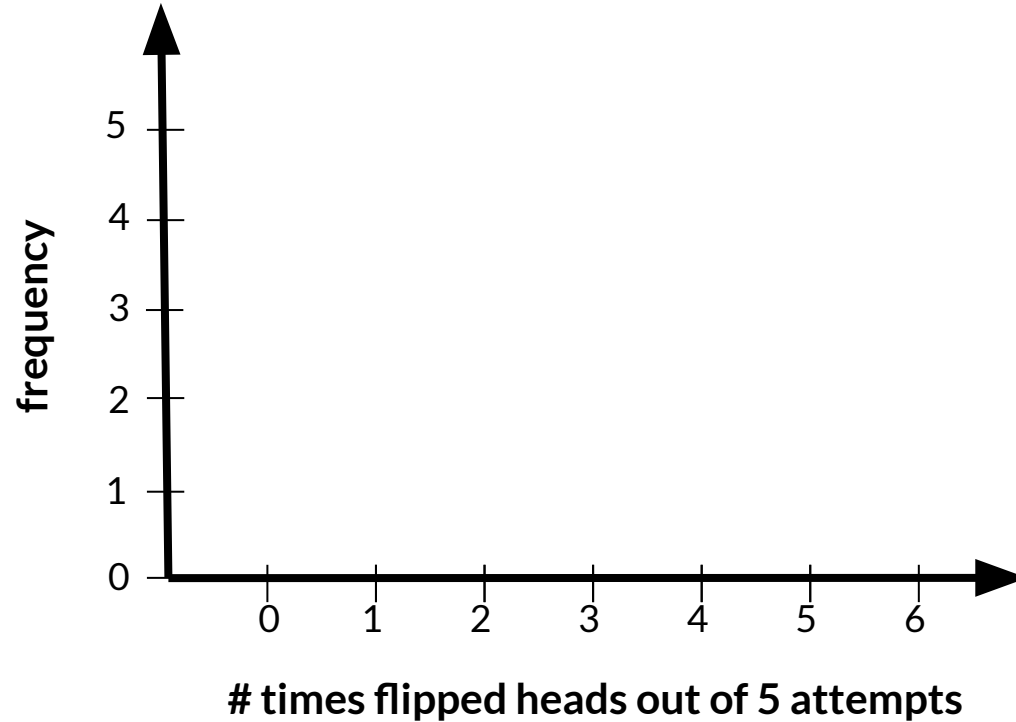
$$Var[X] = N p(1 - p)$$

$$Std[X] = \sqrt{N p(1 - p)}$$

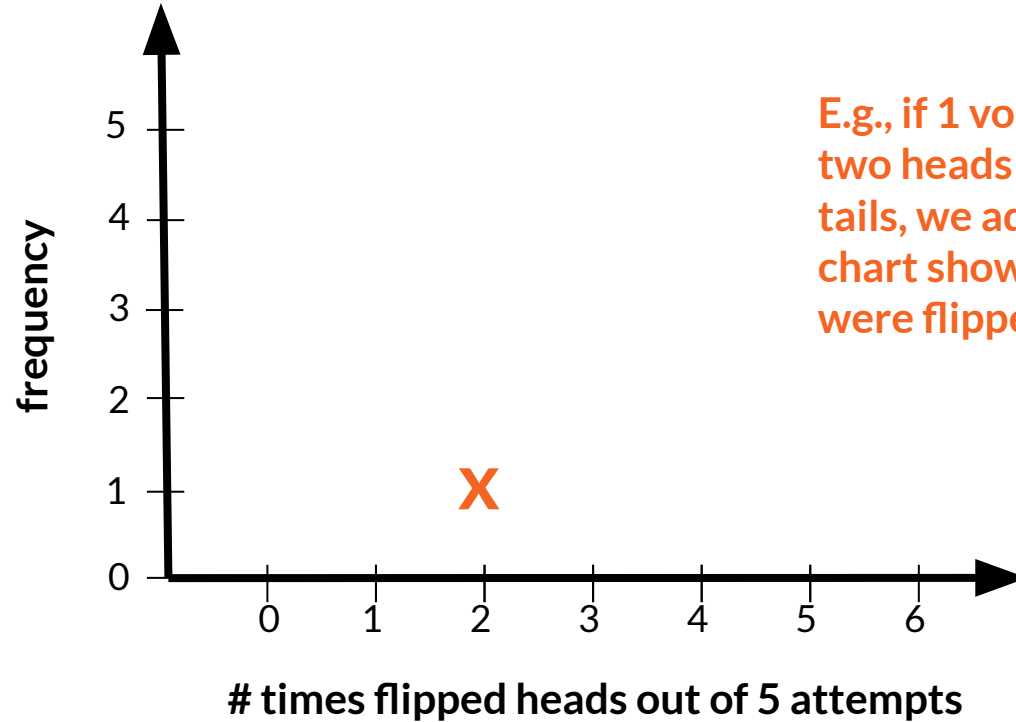
To understand yes/no probability functions...

- We need to understand distributions!
 - We'll build up our understanding starting from histograms
- We need **5 volunteers**

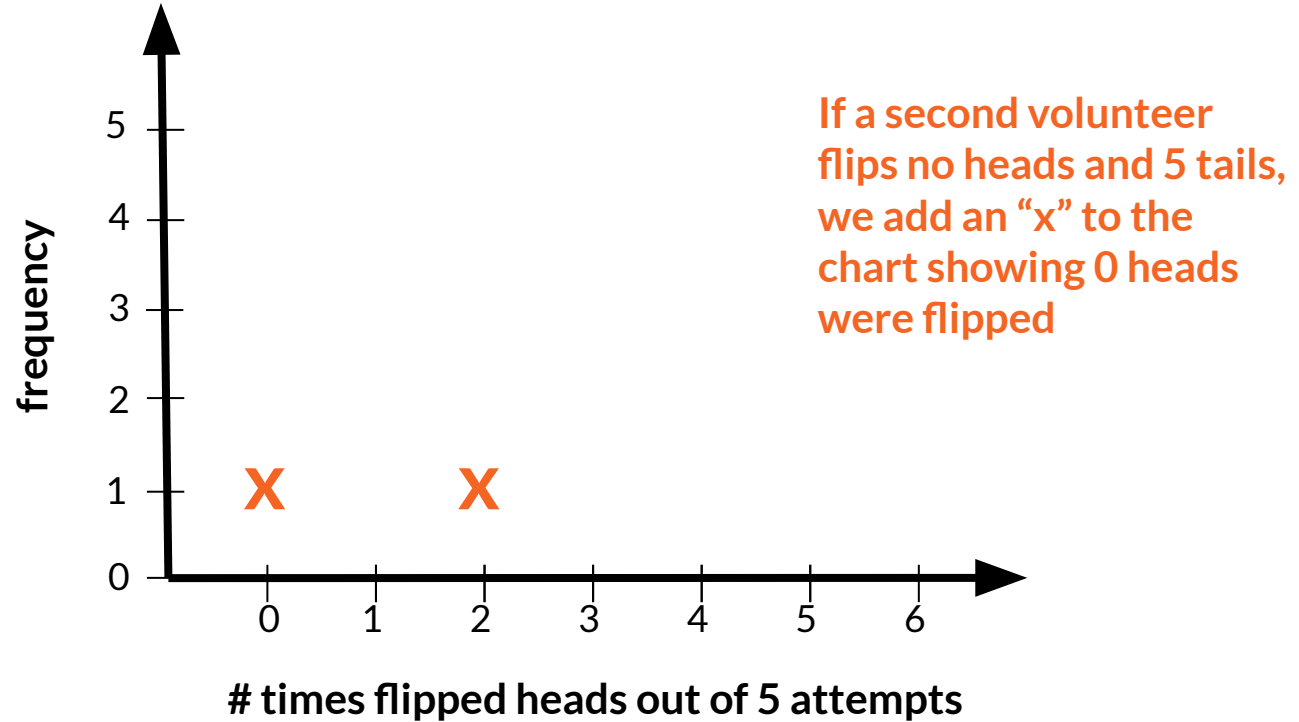
Your job: fill out this chart



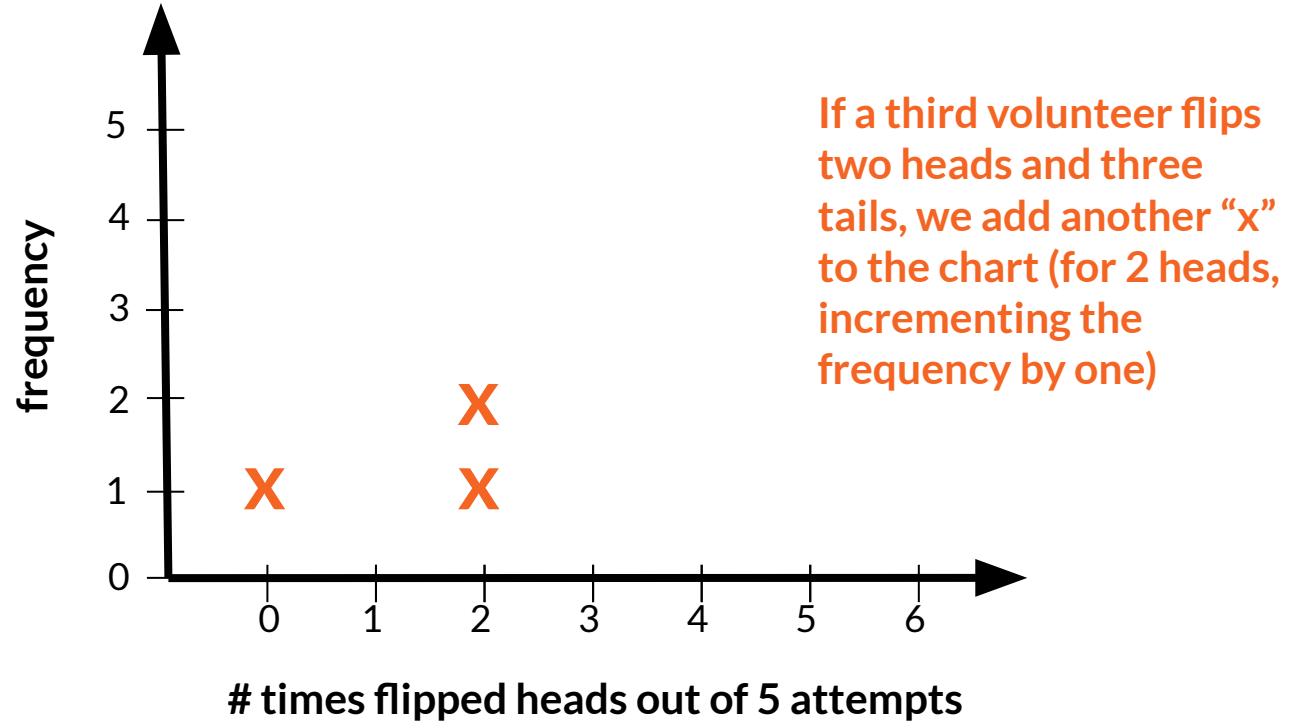
Your job: fill out this chart



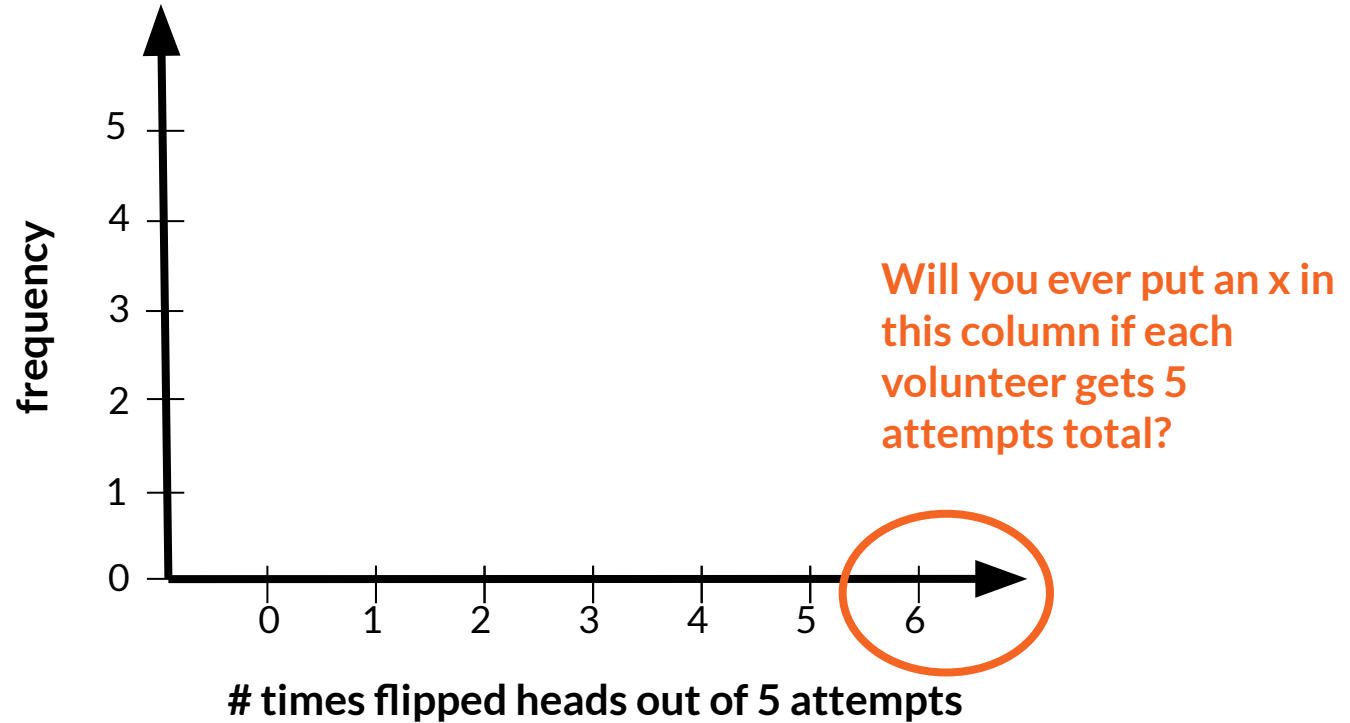
Your job: fill out this chart



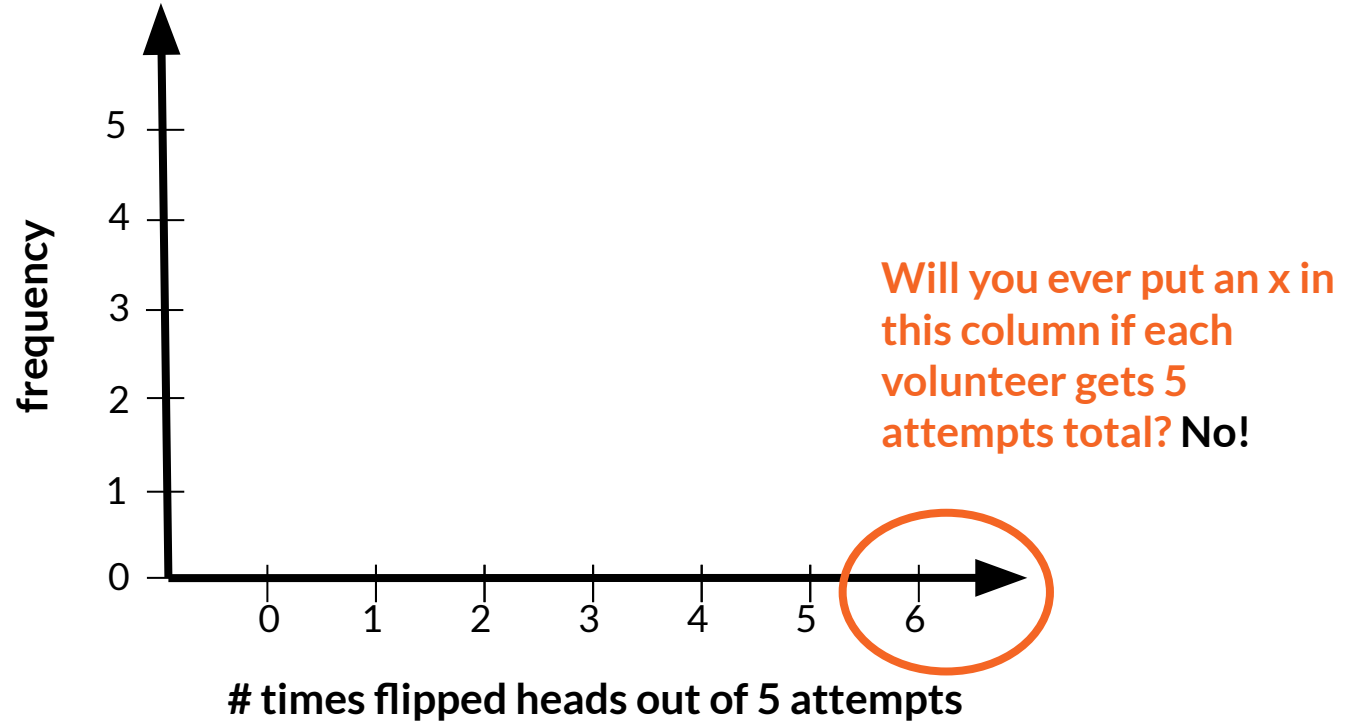
Your job: fill out this chart



Your job: fill out this chart

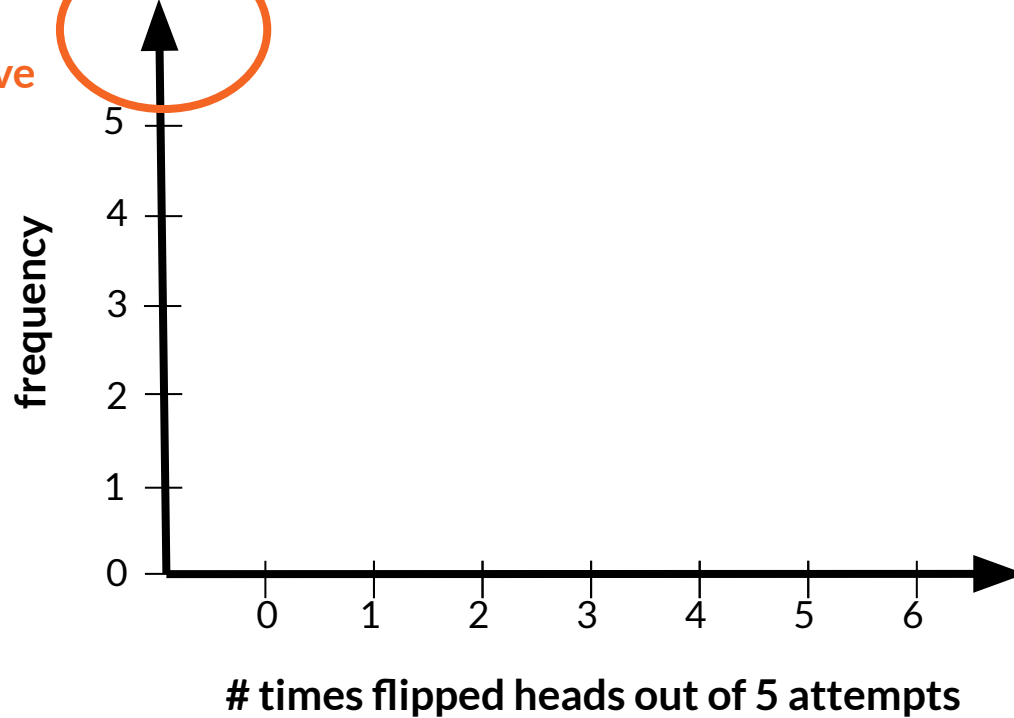


Your job: fill out this chart



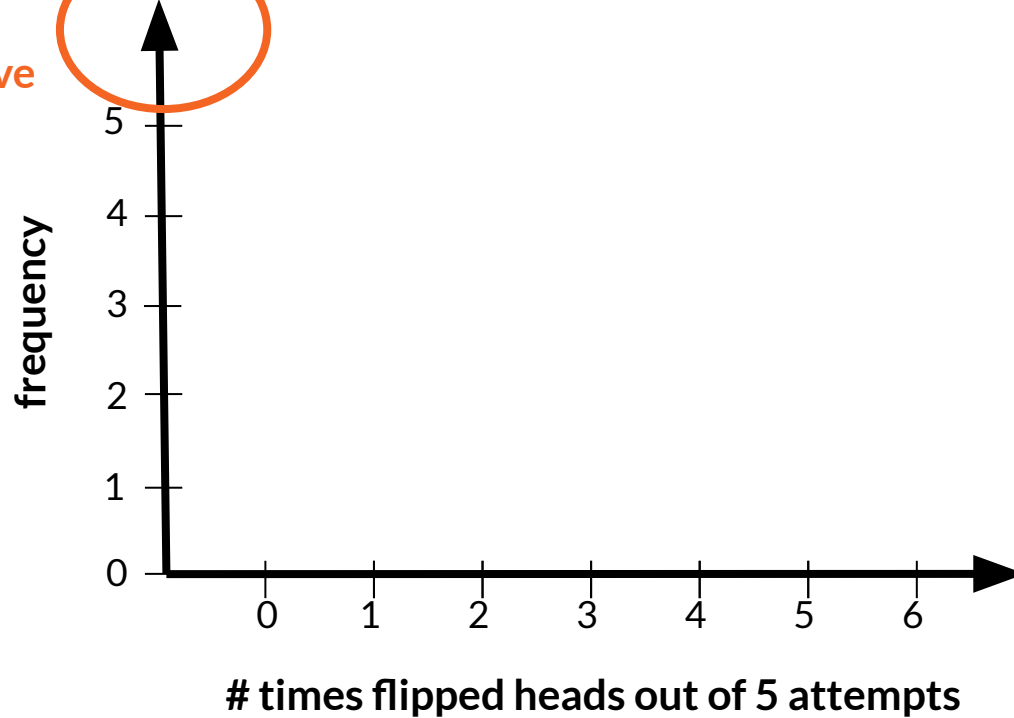
Your job: fill out this chart

Will you ever have
frequency > 5 if we have
5 volunteers flipping
coins?

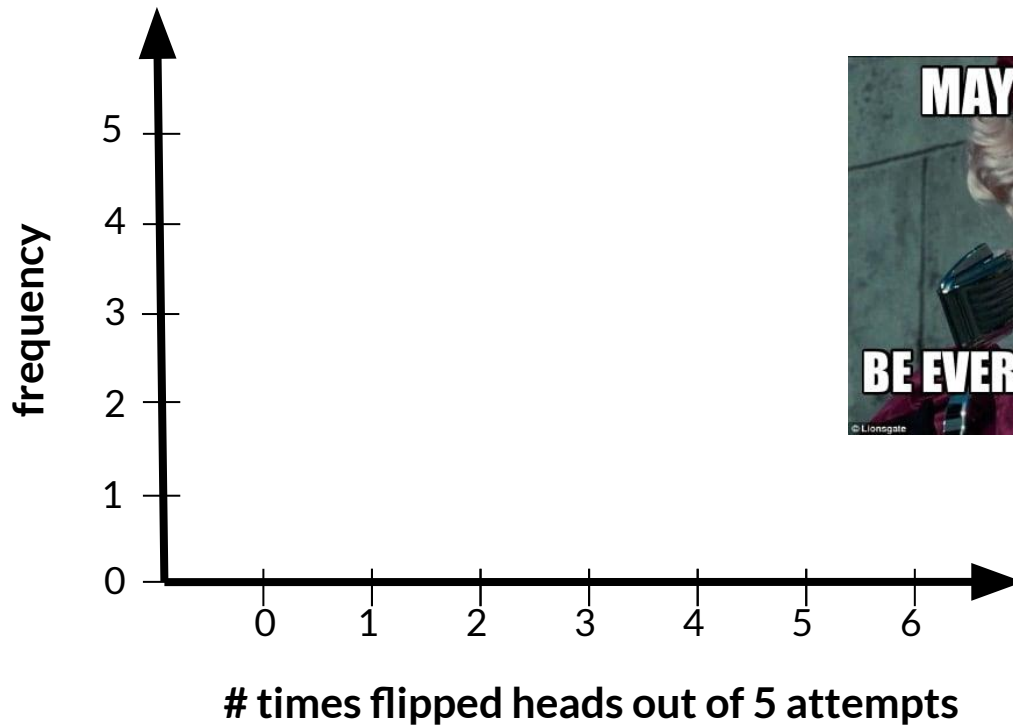


Your job: fill out this chart

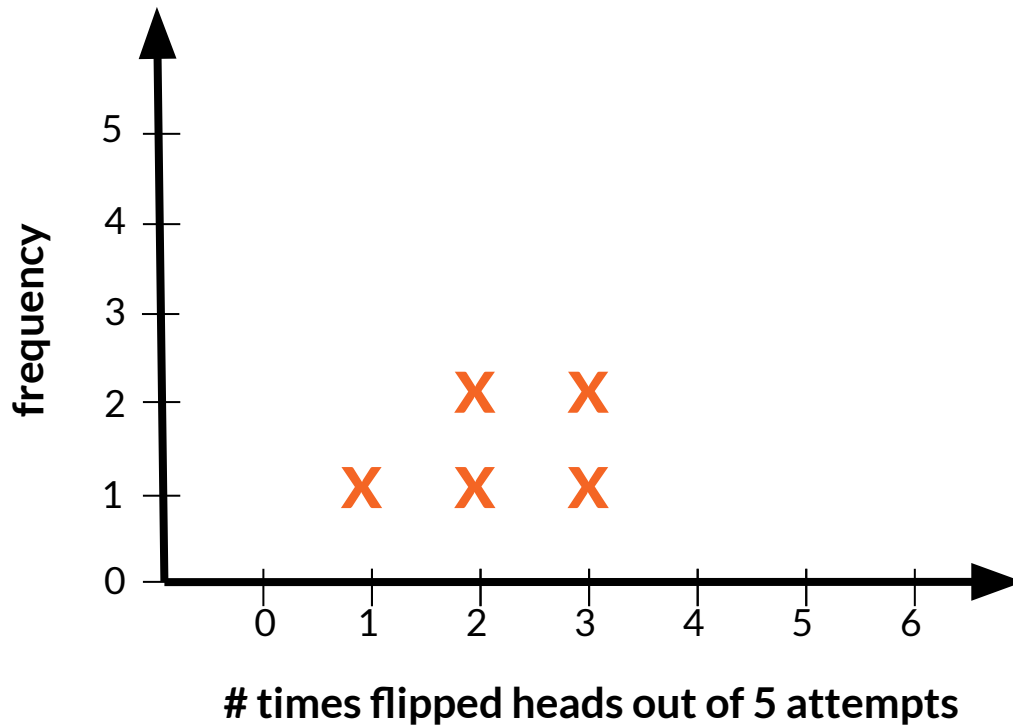
Will you ever have
frequency > 5 if we have
5 volunteers flipping
coins? No!



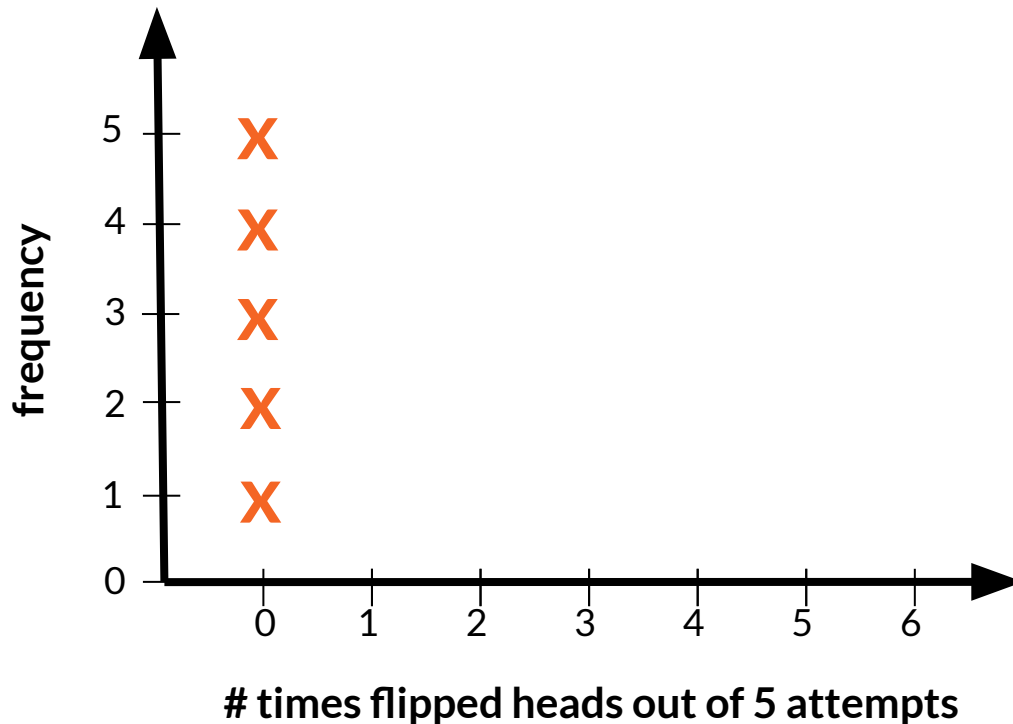
Let the games begin!



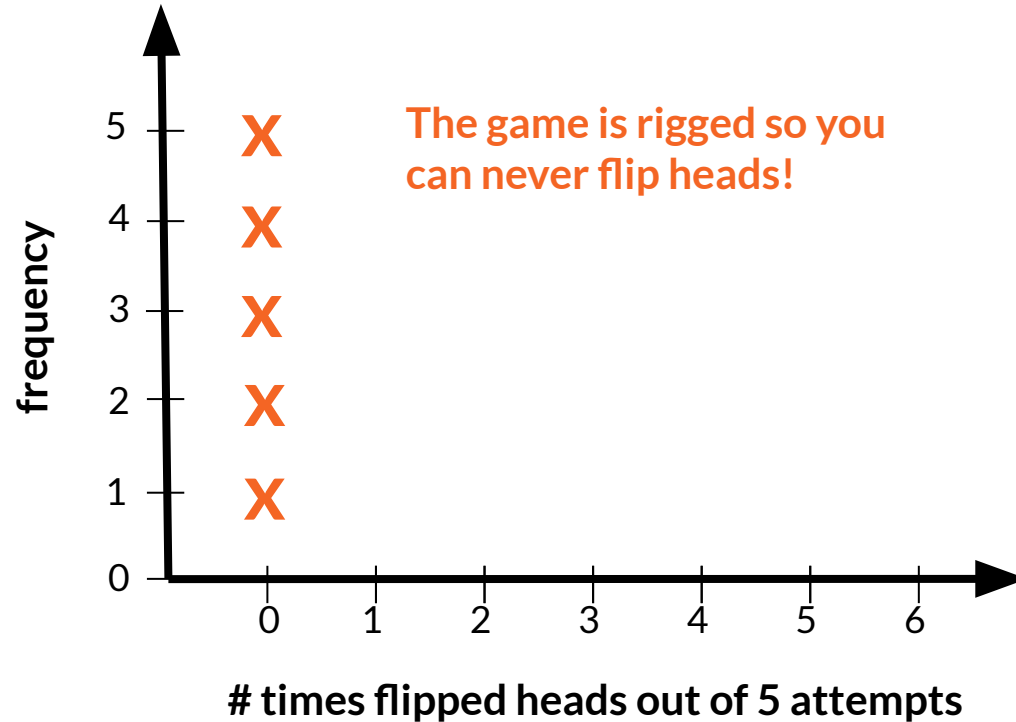
Here's our guess before class:



If this were to happen, what would we think?



If this were to happen, what would we think?



Digression: are coin flips fair?

SIAM REVIEW
Vol. 49, No. 2, pp. 211–235

© 2007 Society for Industrial and Applied Mathematics

Dynamical Bias in the Coin Toss*

Persi Diaconis[†]

Susan Holmes[‡]

Richard Montgomery[§]

Is p actually 0.5 with coins: **yes or no?**

Digression: are coin flips fair?

SIAM REVIEW
Vol. 49, No. 2, pp. 211–235

© 2007 Society for Industrial and Applied Mathematics

Dynamical Bias in the Coin Toss*

Persi Diaconis[†]

Susan Holmes[‡]

Richard Montgomery[§]

Abstract. We analyze the natural process of flipping a coin which is caught in the hand. We show that vigorously flipped coins tend to come up the same way they started. The limiting chance of coming up this way depends on a single parameter, the angle between the normal to the coin and the angular momentum vector. Measurements of this parameter based on high-speed photography are reported. For natural flips, the chance of coming up as started is about .51.

Key words. Berry phase, randomness, precession, image analysis

Digression: are coin flips fair?

arXiv > math > arXiv:2310.04153

Search...

Help | Advanced S

Mathematics > History and Overview

[Submitted on 6 Oct 2023 (v1), last revised 10 Oct 2023 (this version, v2)]

Fair coins tend to land on the same side they started: Evidence from 350,757 flips

František Bartoš, Alexandra Sarafoglou, Henrik R. Godmann, Amir Sahrani, David Klein Leunk, Pierre Y. Gui, David Voss, Kaleem Ullah, Malte J. Zoubek, Franziska Nippold, Frederik Aust, Felipe F. Vieira, Chris-Gabriel Islam, Anton J. Zoubek, Sara Shabani, Jonas Petter, Ingeborg B. Roos, Adam Finnemann, Aaron B. Lob, Madlen F. Hoffstadt, Jason Nak, Jill de Ron, Koen Derks, Karoline Huth, Sjoerd Terpstra, Thomas Bastelica, Magda Matetovici, Vincent L. Ott, Andreea S. Zetea, Katharina Karnbach, Michelle C. Donzallaz, Arne John, Roy M. Moore, Franziska Assion, Riet van Bork, Theresa E. Leiding, Xiaochang Zhao, Adrian Karami Motaghi, Ting Pan, Hannah Armstrong, Tianqi Peng, Mara Bialas, Joyce Y.-C. Pang, Bohan Fu, Shujun Yang, Xiaoyi Lin, Dana Sleiffer, Miklos Bognar, Balazs Aczel, Eric-Jan Wagenmakers

Many people have flipped coins but few have stopped to ponder the statistical and physical intricacies of the process. In a preregistered study we collected 350,757 coin flips to test the counterintuitive prediction from a physics model of human coin tossing developed by Diaconis, Holmes, and Montgomery (D-H-M; 2007). The model asserts that when people flip an ordinary coin, it tends to land on the same side it started -- D-H-M estimated the probability of a same-side outcome to be about 51%. Our data lend strong support to this precise prediction: the coins landed on the same side more often than not, $\Pr(\text{same side}) = 0.508$, 95% credible interval (CI) [0.506, 0.509], $\text{BF}_{\text{same-side bias}} = 2364$. Furthermore, the data revealed considerable between-people variation in the degree of this same-side bias. Our data also confirmed the generic prediction that when people flip an ordinary coin -- with the initial side-up randomly determined -- it is equally likely to land heads or tails: $\Pr(\text{heads}) = 0.500$, 95% CI [0.498, 0.502], $\text{BF}_{\text{heads-tails bias}} = 0.183$. Furthermore, this lack of heads-tails bias does not appear to vary across coins. Our data therefore provide strong evidence that when some (but not all) people flip a fair coin, it tends to land on the same side it started. Our data provide compelling statistical support for D-H-M physics model of coin tossing.

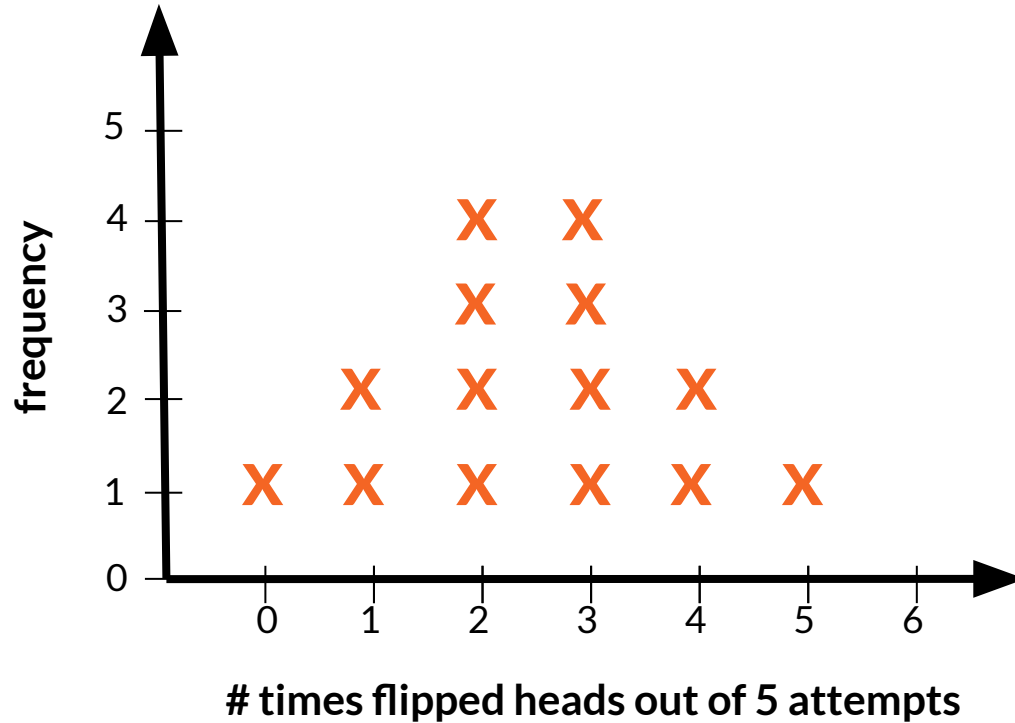
Subjects: **History and Overview** (math.HO); Data Analysis, Statistics and Probability (physics.data-an); Other Statistics (stat.OT)

Cite as: arXiv:2310.04153 [math.HO]

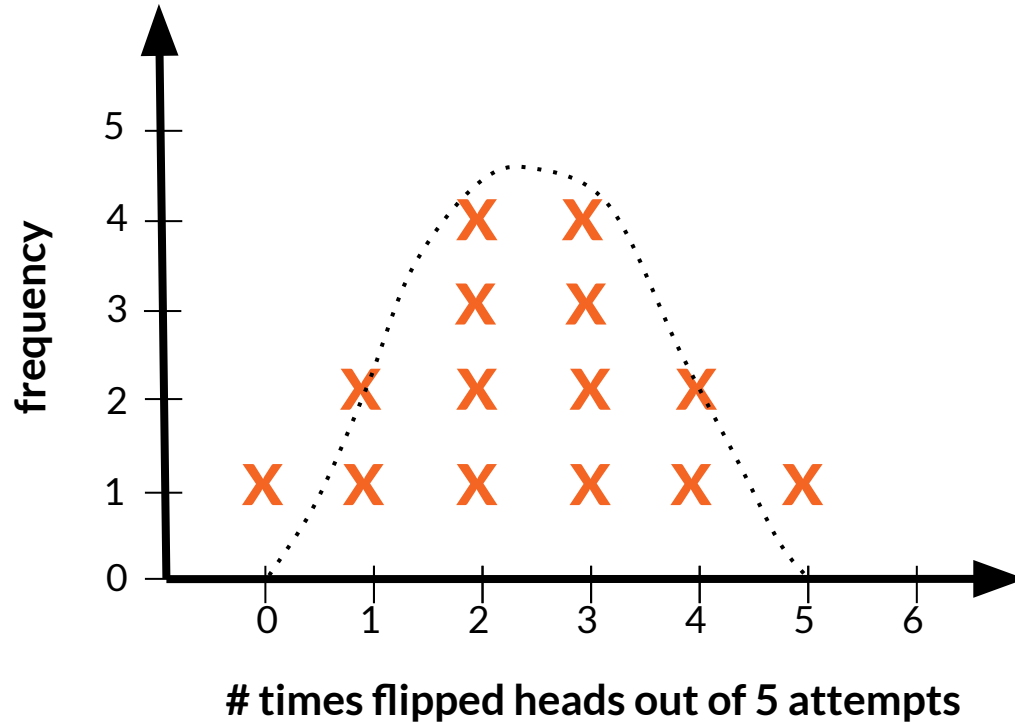
(or arXiv:2310.04153v2 [math.HO] for this version)

<https://doi.org/10.48550/arXiv.2310.04153> 

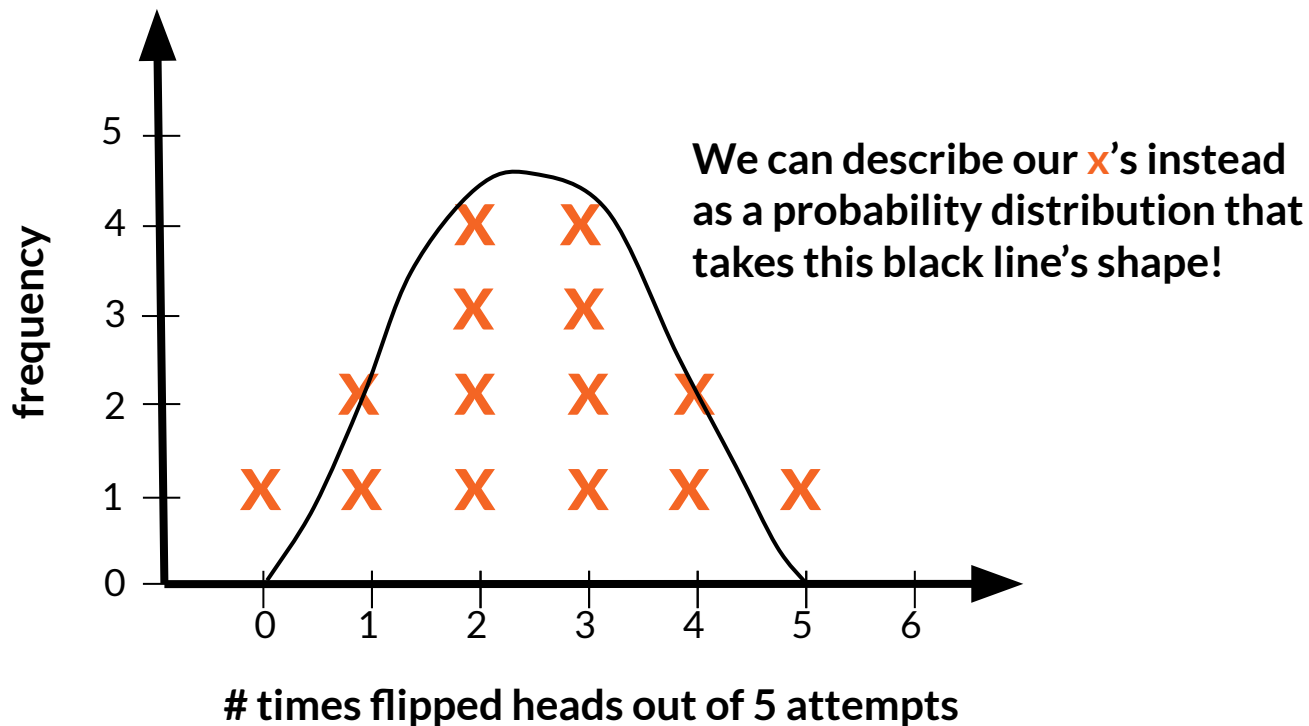
If we had more volunteers, maybe we'd get:



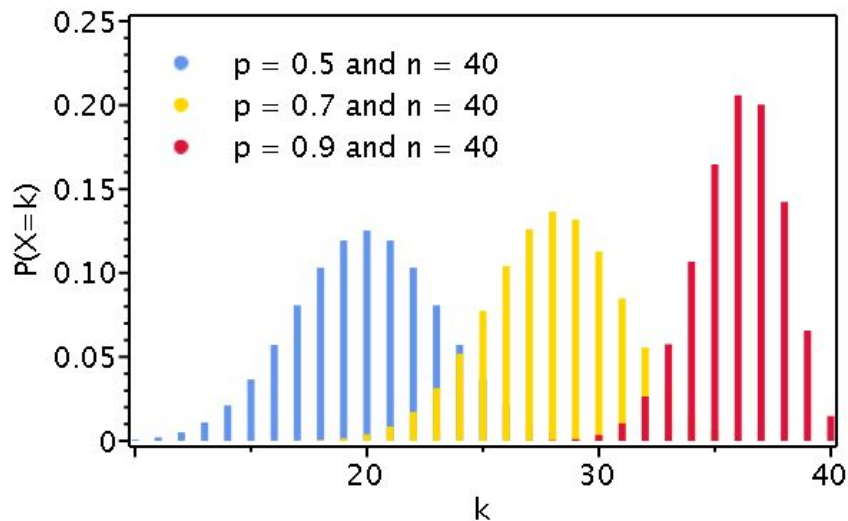
If we had more volunteers, maybe we'd get:



If we had more volunteers, maybe we'd get:

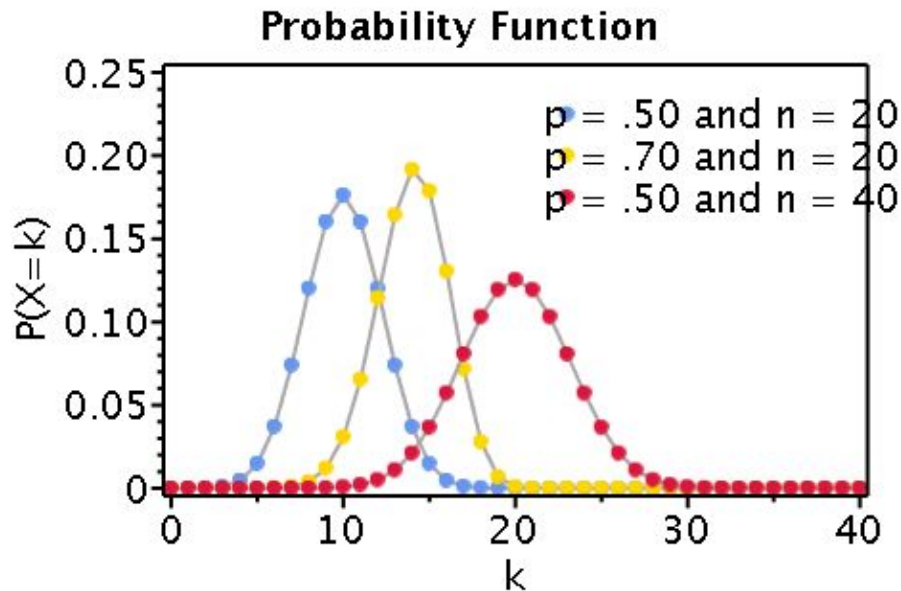
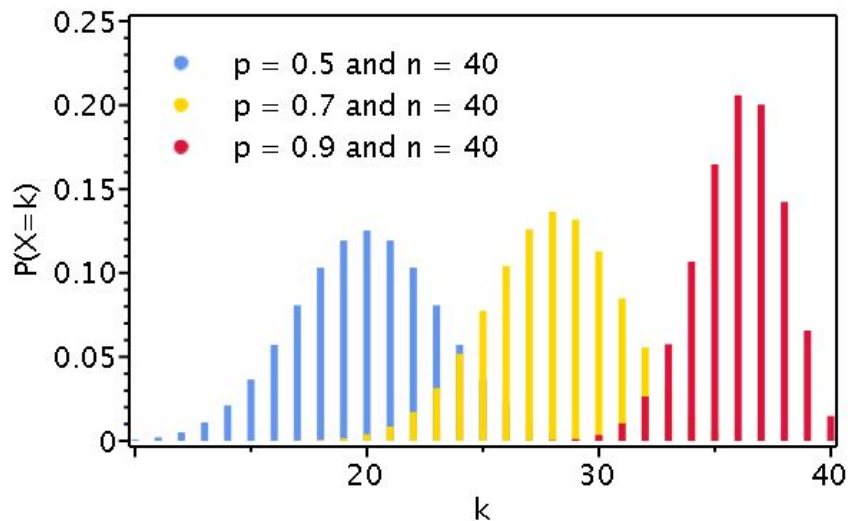


Histograms → Probability Functions



Blue histogram: same $p = 0.5$ (tested on 40 flips instead of 5 flips per volunteer) but the y-axis is now normalized to be probabilities of the counts

Histograms → Probability Functions



(*notice the n values have changed on the right so these distributions don't match up exactly to the left)

How do you know if you can use a binomial distribution?


1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

How do you know if you can use a binomial distribution?

- “Bernoulli trial” →
1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
 2. You conduct each “trial” exactly the same way for a fixed number of times n
 3. The probability of success p is the same for each trial
 4. Trials are independent

How do you know if you can use a binomial distribution?


“repeated Bernoulli trials”: conditions are the same for each trial (identically distributed)

- 
1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
 2. You conduct each “trial” exactly the same way for a fixed number of times n
 3. The probability of success p is the same for each trial
 4. Trials are independent

How do you know if you can use a binomial distribution?

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

the result of one trial
doesn't affect the
result of another trial



Notation

X = number of successes

If our 4 conditions are met, X is a “**binomial random variable**”

We can talk about the probability $P(X=k)$
where k = a specific number of outcomes
we're interested in

Notation

X = number of successes

If our 4 conditions are met, X is a “**binomial random variable**”

E.g., for 10 coin flips with a fair coin, define $P(X=5)$ as the probability that you get 5 heads

We can talk about the probability $P(X=k)$ where k = a specific number of outcomes we're interested in

Notation

X = number of successes

If our 4 conditions are met, X is a “**binomial random variable**”

E.g., for 10 coin flips
with a fair coin, does
 $P(X=11)$ make sense?

We can talk about the probability $P(X=k)$
where k = a specific number of outcomes
we're interested in

Notation

X = number of successes

If our 4 conditions are met, X is a “**binomial random variable**”

E.g., for 10 coin flips
with a fair coin, does
 $P(X=11)$ make sense?
Not really, but it'll be 0.

We can talk about the probability $P(X=k)$
where k = a specific number of outcomes
we're interested in

Is X a binomial random variable?

X = number of successes

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

Toss a weighted coin 100 times.
 X is the number of heads.
Probability of heads is 70% each time.

Is X a binomial random variable?

X = **number of successes**

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

Toss a weighted coin 100 times.

X is **the number of heads**.

Probability of heads is 70% each time.

Yes, assuming no magic involved

Can you use a binomial distribution?

X = *number of successes*

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

Sample students until you’ve found one who likes cats.

X = number of students sampled.

Probability of a student liking cats is 28%.

Can you use a binomial distribution?

X = number of successes

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for **a fixed number of times n**
3. The probability of success p is the same for each trial
4. Trials are independent

Nope, no n defined!

Sample students until you’ve found one who likes cats.

X = number of students sampled.

Probability of a student liking cats is 28%.

Can you use a binomial distribution?

X = *number of successes*

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

Pumpkin quality is either “rotten” or “acceptable,” with 20% of pumpkins being rotten.

Inspect the quality of 15 pumpkins by randomly sampling 5 pumpkins without replacement.

X = number of acceptable pumpkins.

Can you use a binomial distribution?

X = *number of successes*

1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. **The probability of success p is the same for each trial**
4. Trials are independent

Pumpkin quality is either “rotten” or “acceptable,” with 20% of pumpkins being rotten.

Inspect the quality of 15 pumpkins by randomly sampling 5 pumpkins **without** replacement.

X = number of acceptable pumpkins.


No! First trial: $p = 12/15$. If a rotten pumpkin is selected, the second trial's $p = 12/14$, but if an acceptable pumpkin is selected, $p = 11/14$.

Always check whether you can use a binomial distribution!


1. Each trial has only 2 possible outcomes: “success” (1) or “failure” (0)
2. You conduct each “trial” exactly the same way for a fixed number of times n
3. The probability of success p is the same for each trial
4. Trials are independent

Counting i.i.d. events

If the event is 0 or 1, and we know $p=P(1)$, N = the total number of trials, and X = the number of 1's:

$$\frac{N!}{X!(N-X)!}$$



How many sequences?
(Binomial coefficient)

$$p^X (1-p)^{N-X}$$


Probability of one sequence

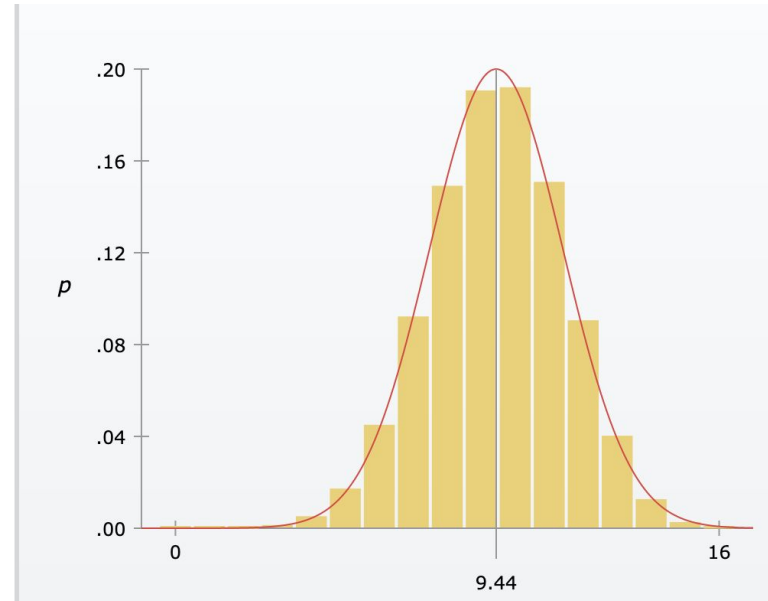
Counting i.i.d. events

If the event is 0 or 1, and we know $p=P(1)$, N = the total number of trials, and X = the number of 1's:

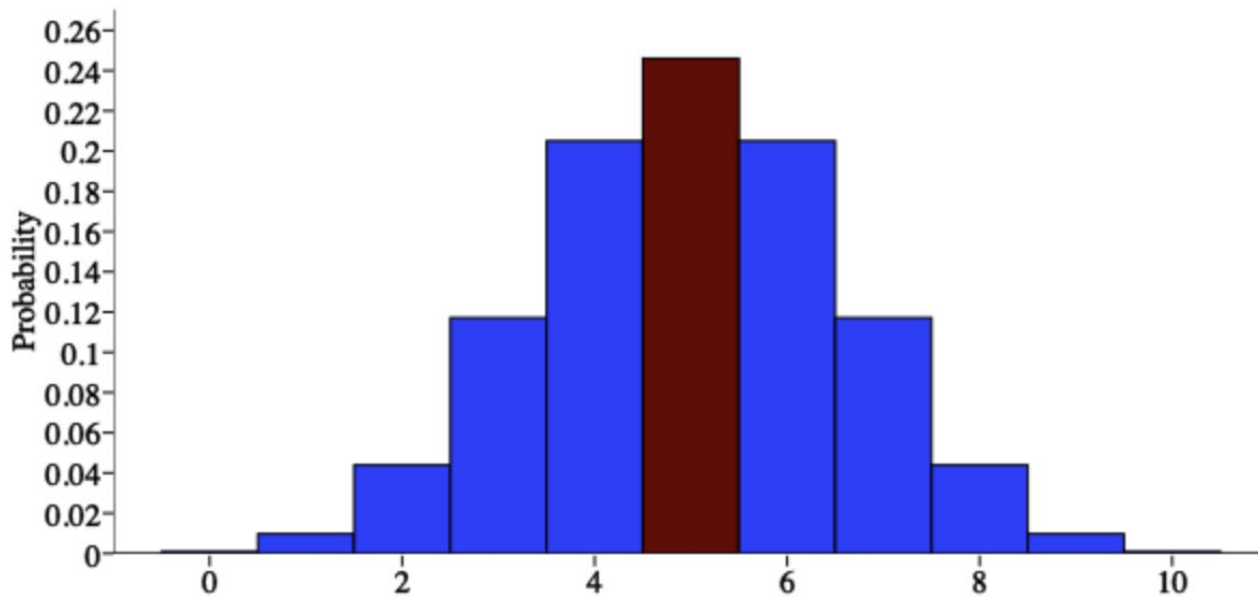
$$\frac{N!}{X!(N-X)!} p^X (1-p)^{N-X}$$


“probability mass function” $P(X)$

Widget: how do n and p affect the distribution?



n = p =



Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

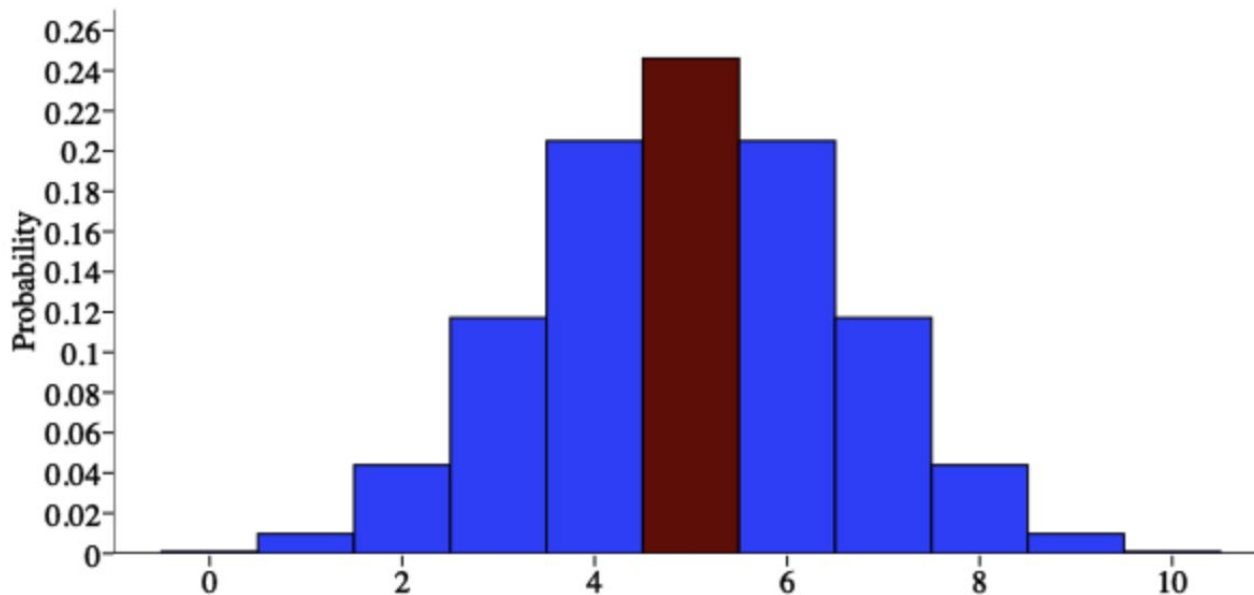
Probability = 0.2461

n = 10

p = 0.5

Plot distribution

Parameters for binomial distribution



Export this graph

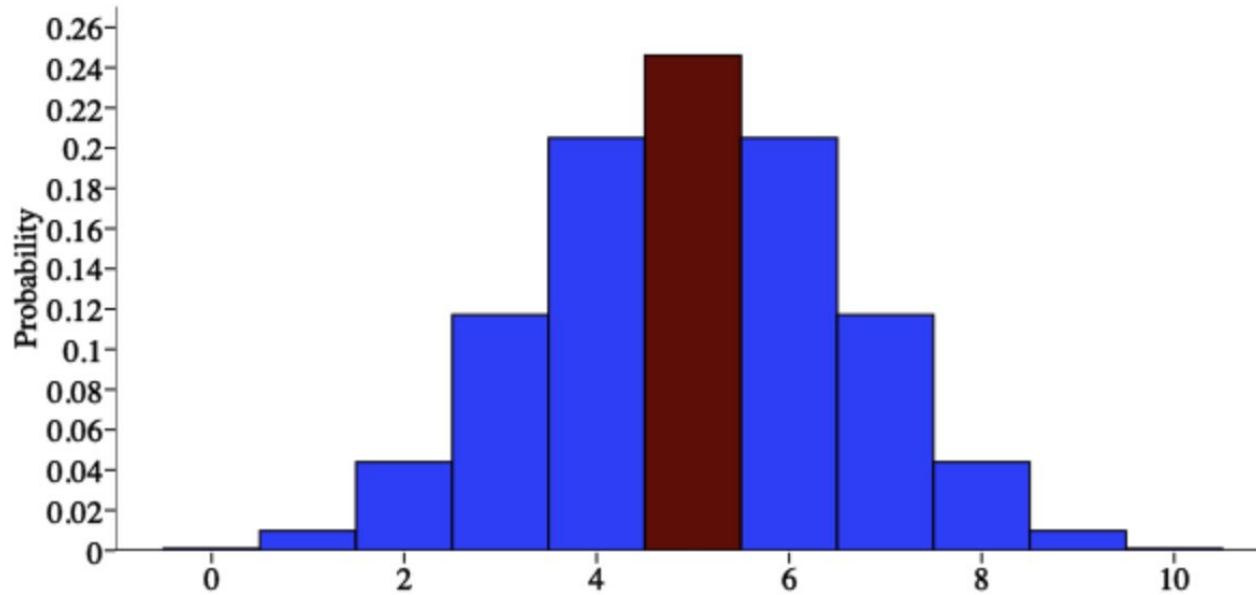
Calculate the probability of exactly 5 successes. Go!

-OR-

Calculate the probability of between and successes (inclusive). Go!

Probability = 0.2461

n = p =



How many combos of exactly 5 heads * pr(getting 5 head sequence)

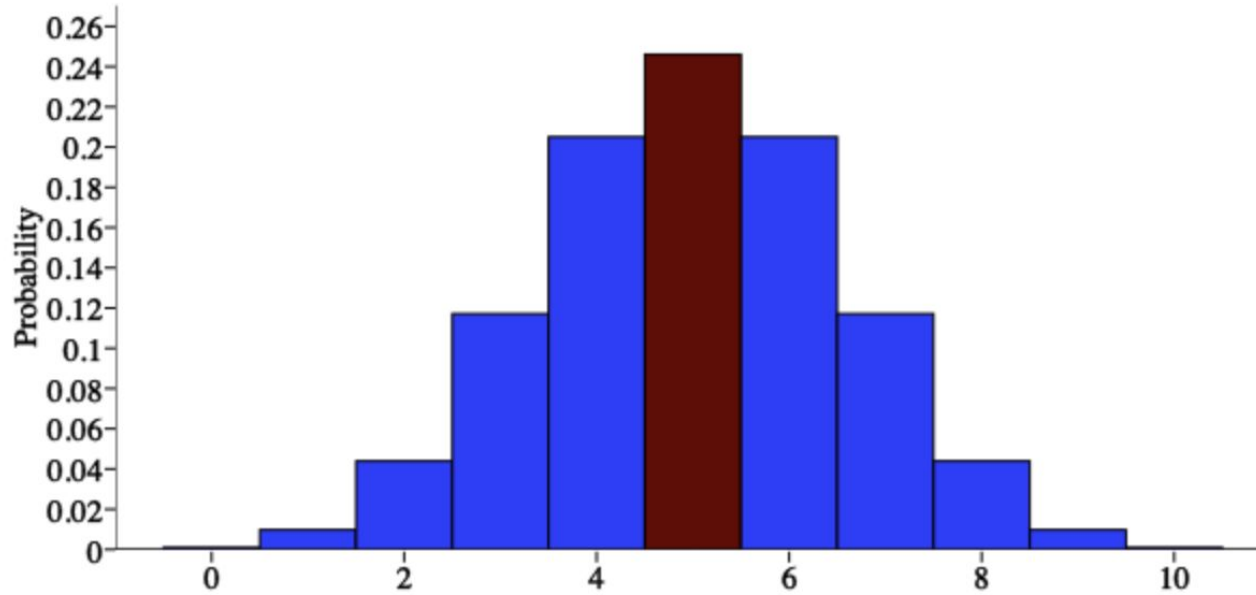
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2461

n = p =



Calculate the probability of successes.

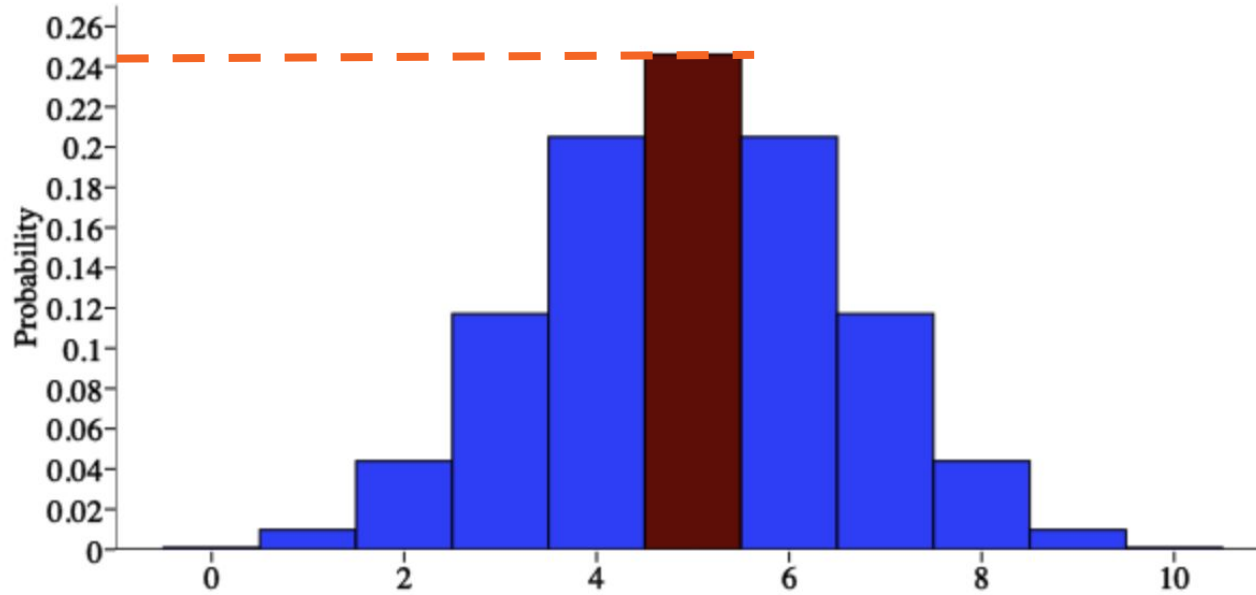
P(X=5)

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2461

n = p =



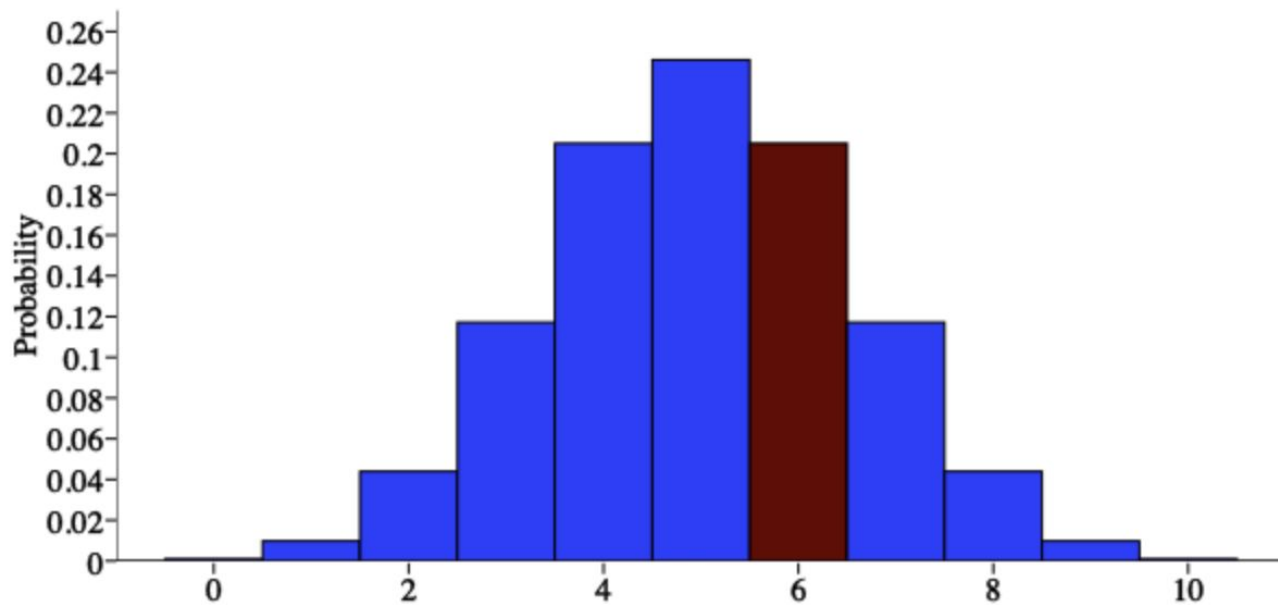
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2461

n = p =



What if we change from wanting 5 heads to 6?

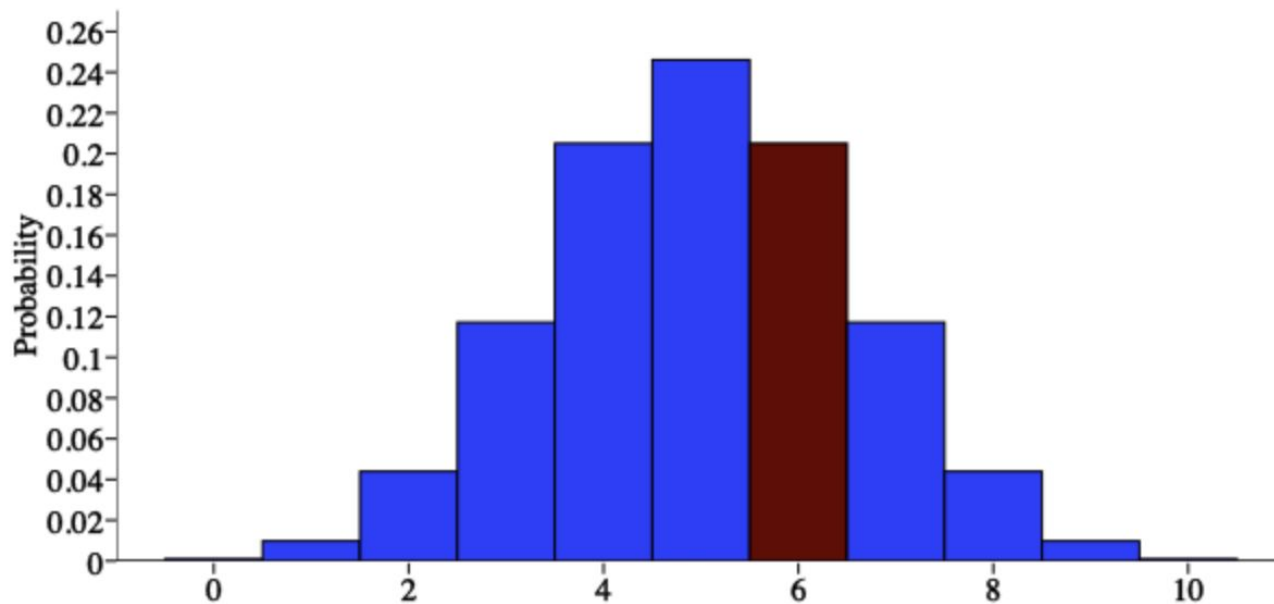
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2051

n = p =



Calculate the probability of successes.

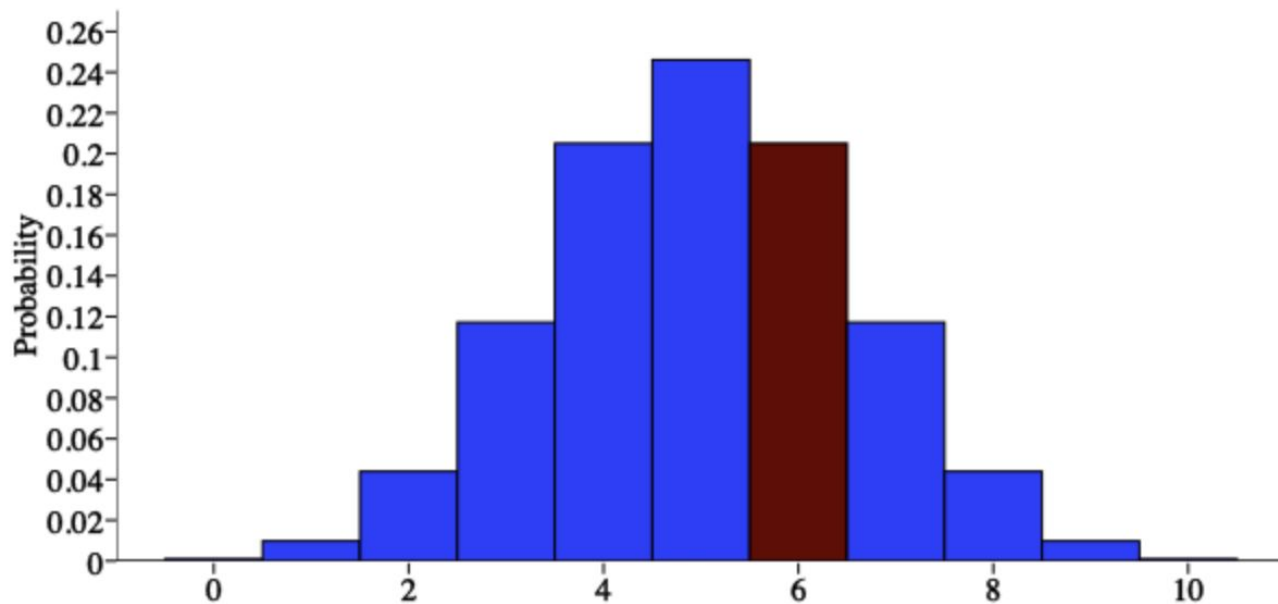
P(?)

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2051

n = p =



Calculate the probability of successes.

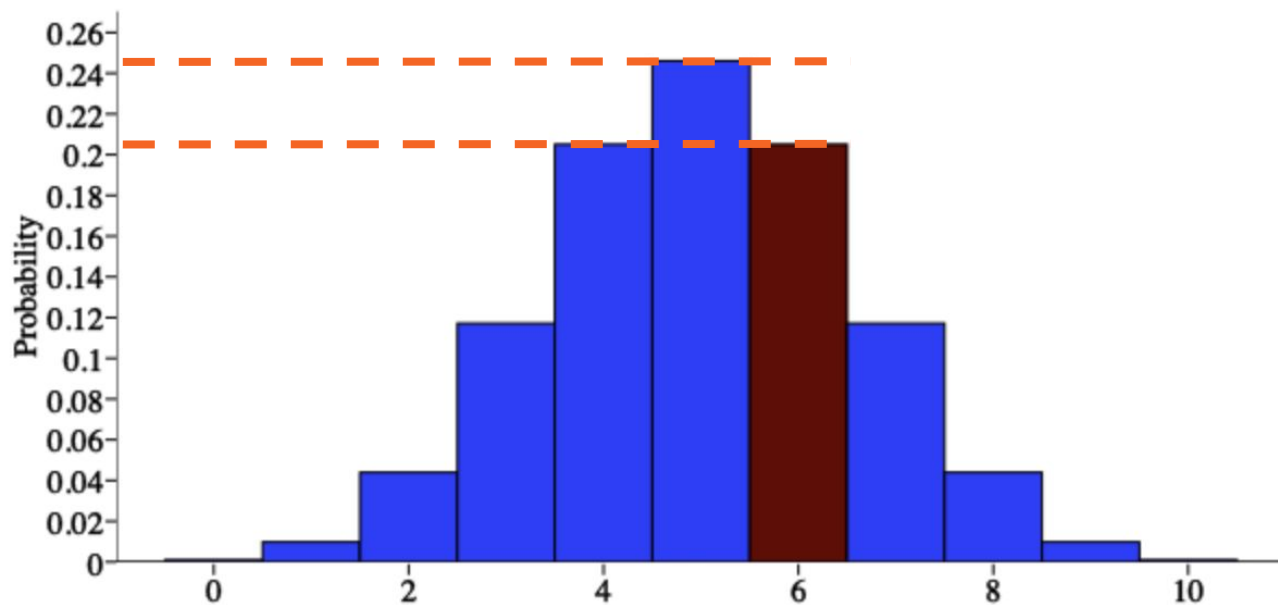
$P(X=6)$

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2051

n = p =



Probability (red bar) is lower than for exactly 5 successes

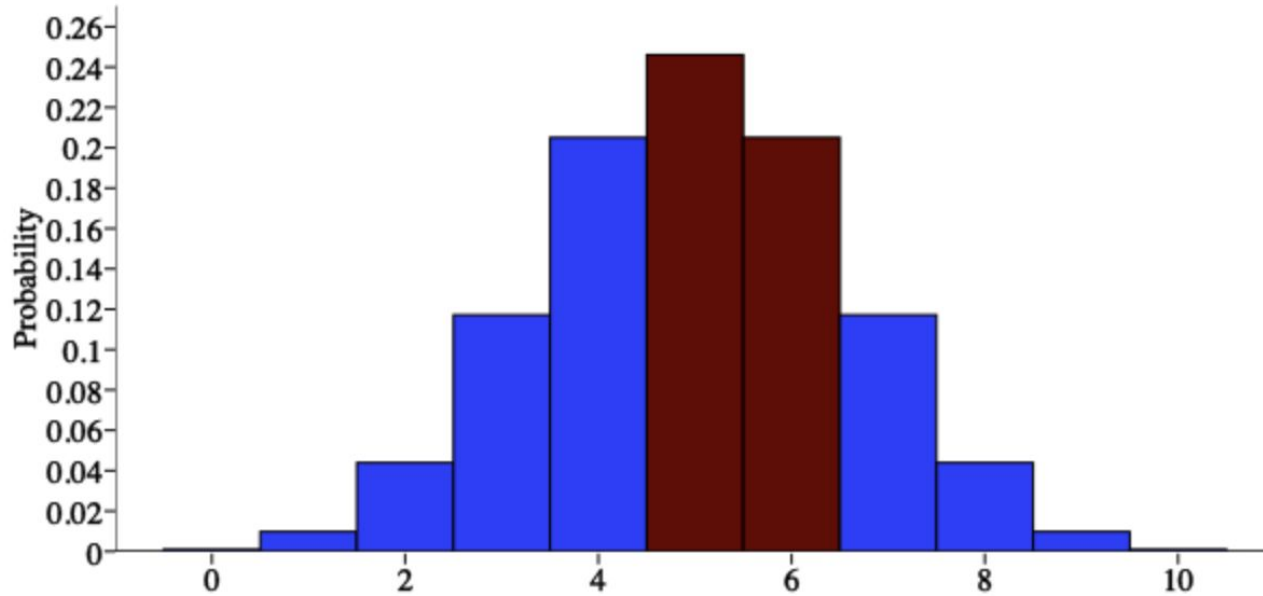
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.2051

n = p =



We can calculate the probability among multiple success counts

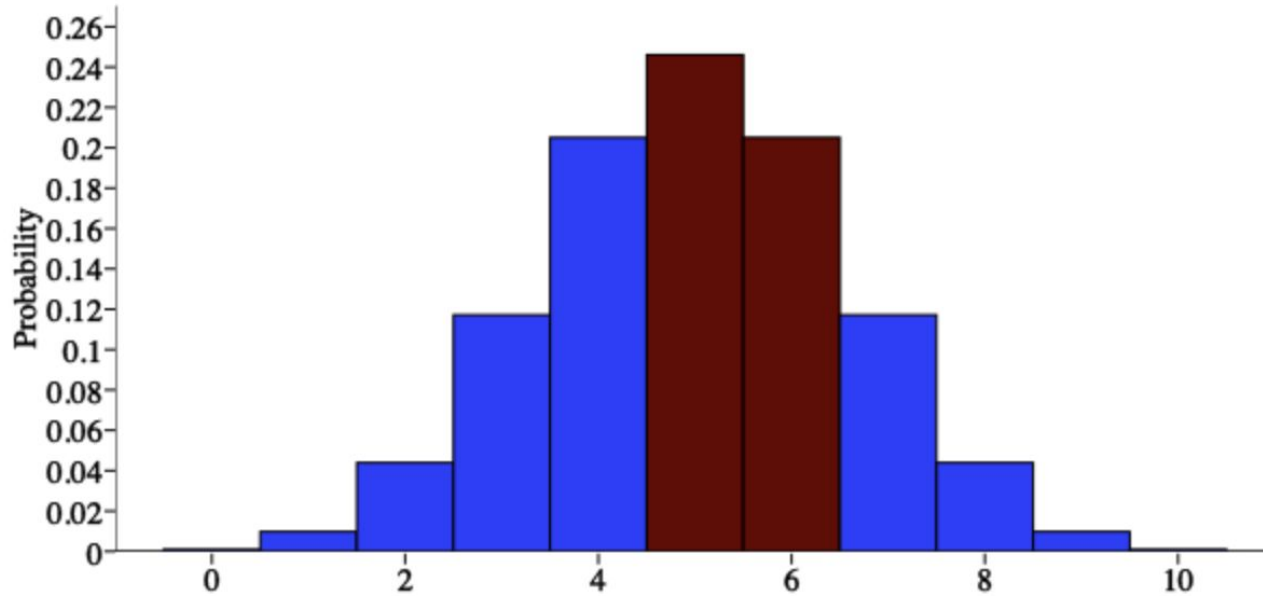
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.4512

n = p =



Calculate the probability of successes.

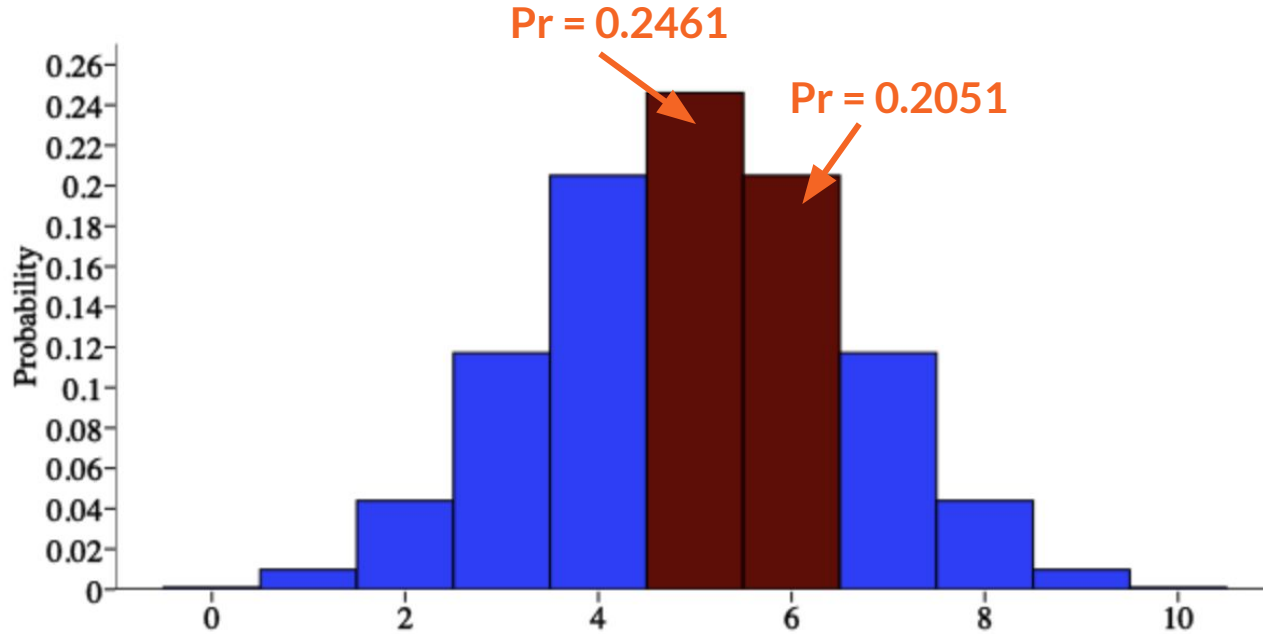
$P(5 \leq X \leq 6)$

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.4512

n = p =



We can calculate the probability among multiple success counts
... by summing each of the red bars

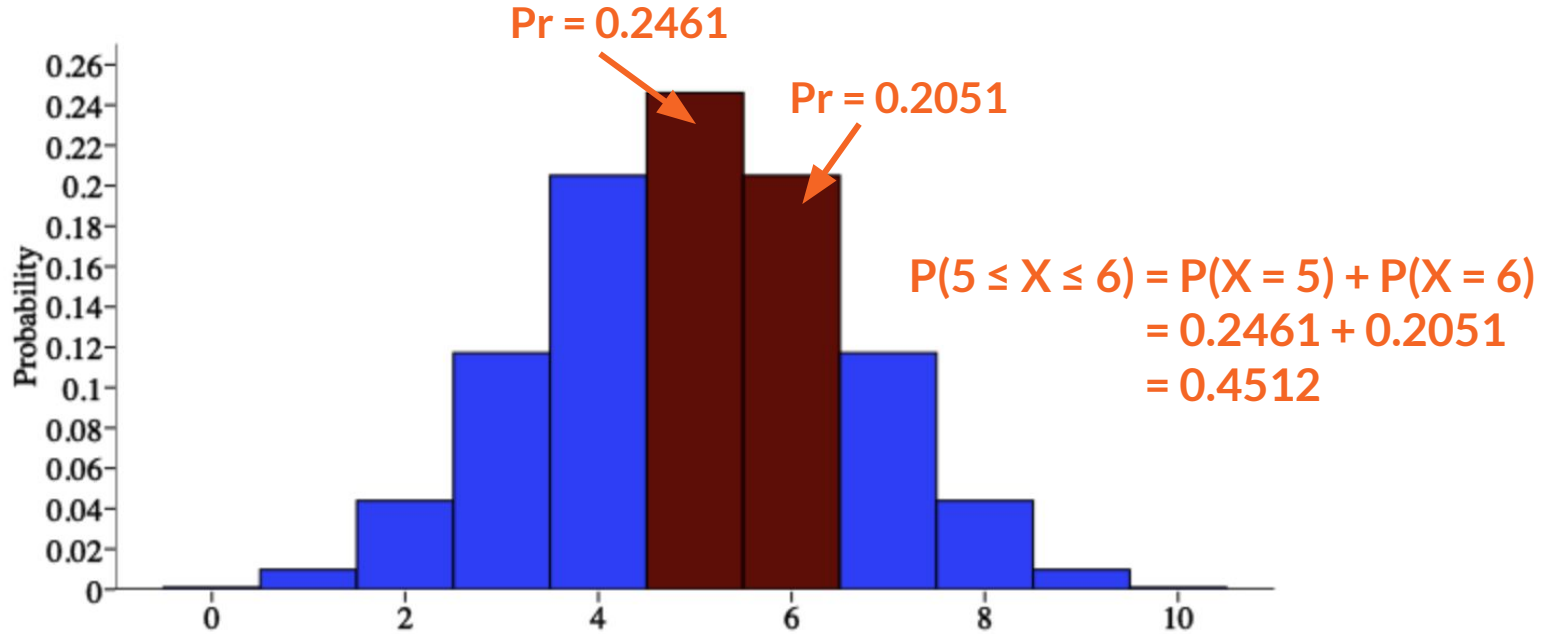
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.4512

n = p =



We can calculate the probability among multiple success counts
... by summing each of the red bars

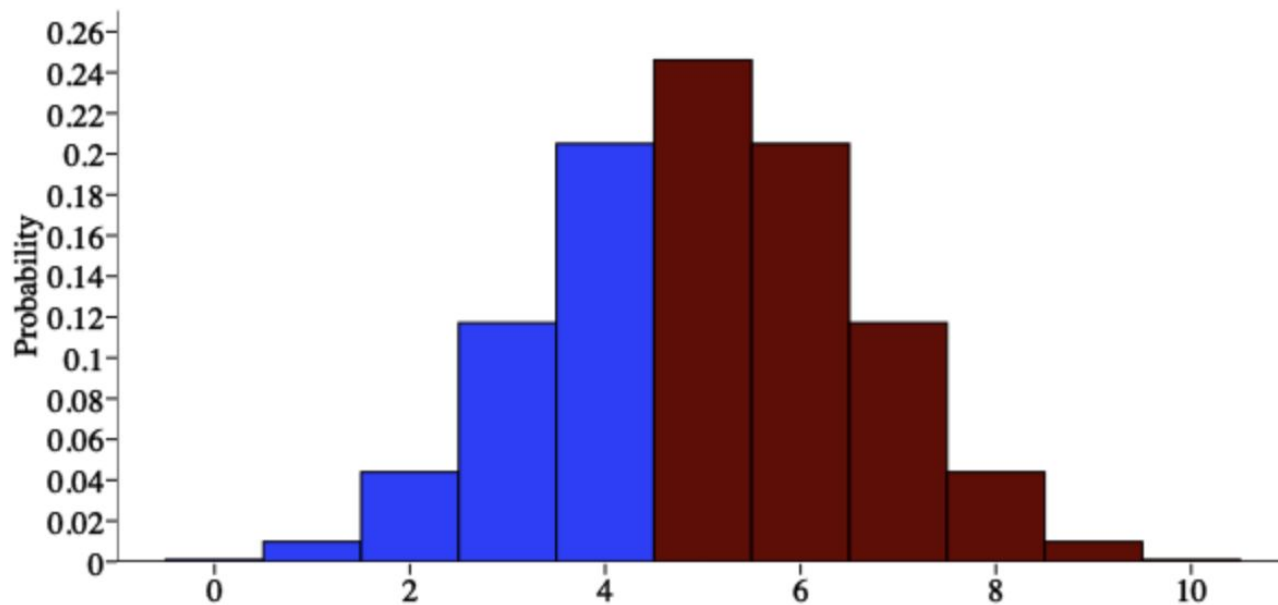
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.4512

n = p =



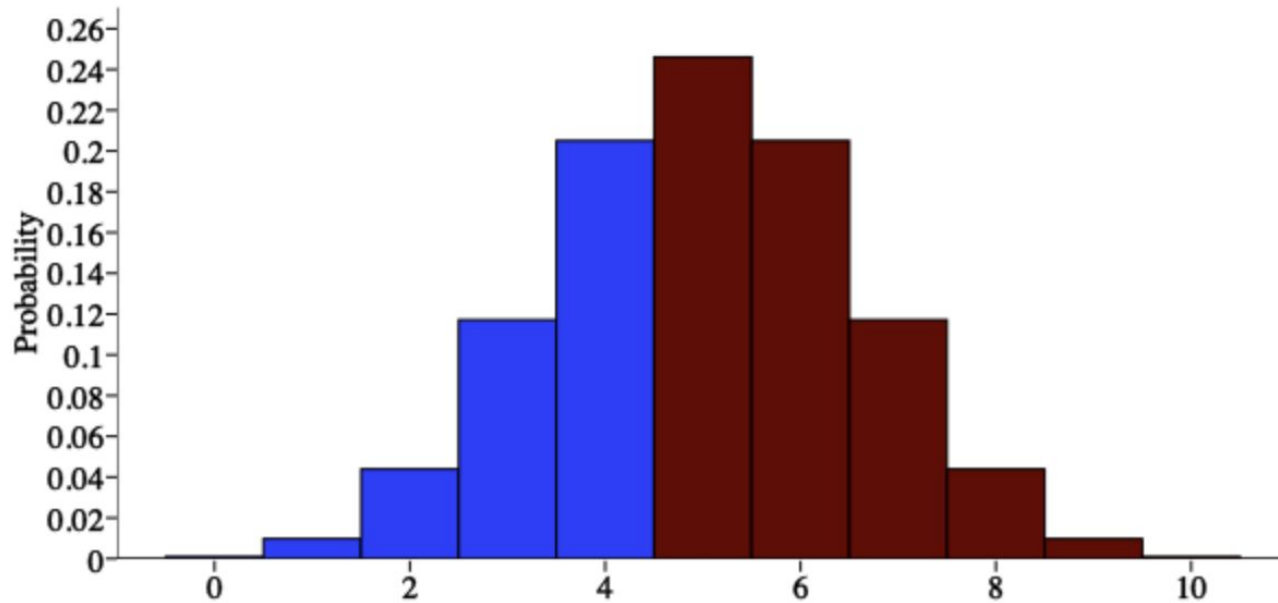
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.623

n = p =



Calculate the probability of successes.

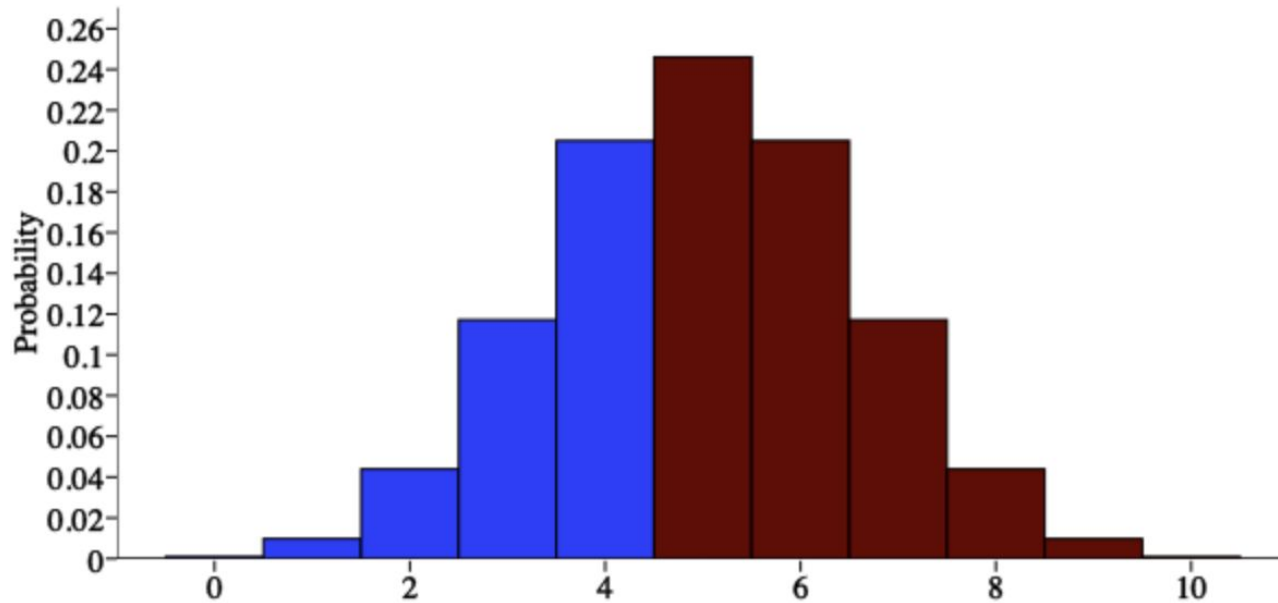
$P(X \geq 5)$

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.623

n = p =



This has a special name: the **cumulative binomial probability**

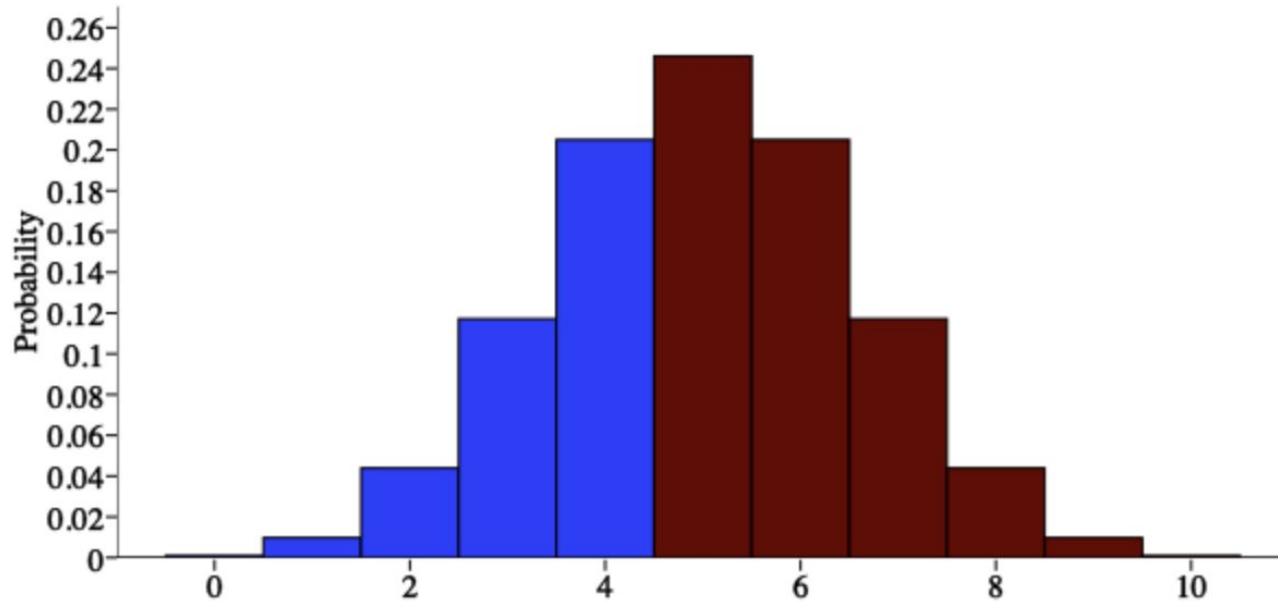
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.623

n = p =



This is equivalent to calculating the probability between 5 and 10 successes inclusive (since our max # coin flips is 10)

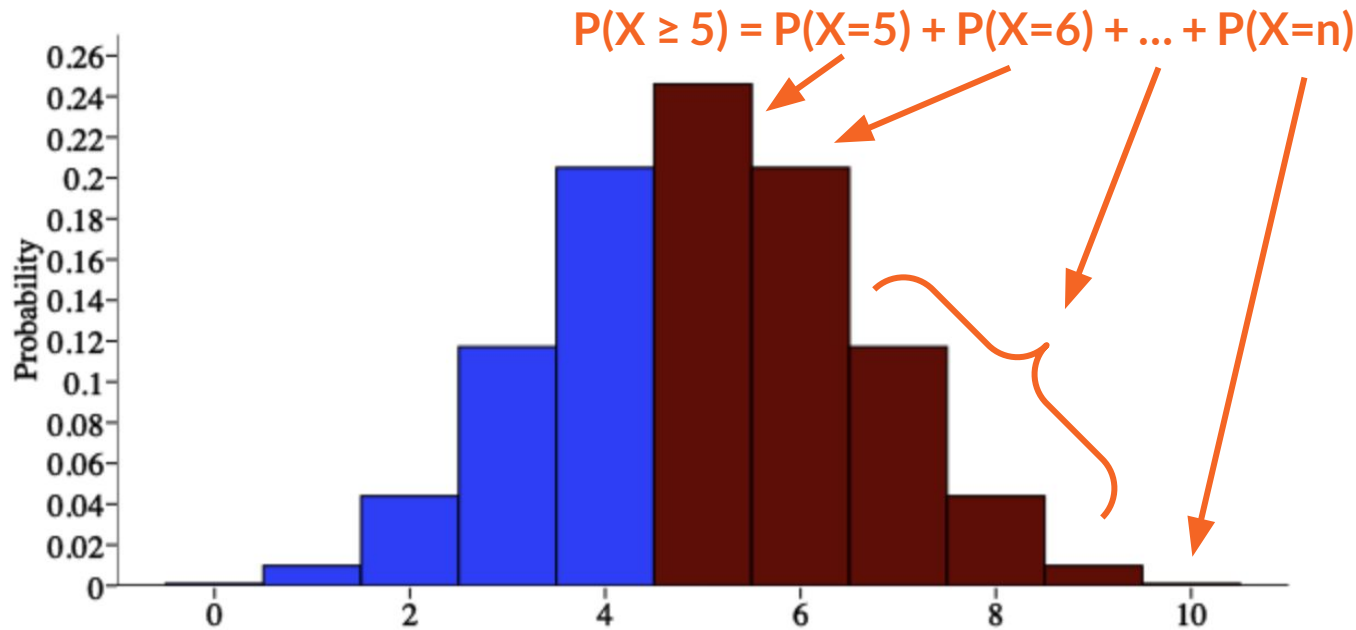
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.623

n = p =



This is equivalent to calculating the probability between 5 and 10 successes inclusive (since our max # coin flips is 10)

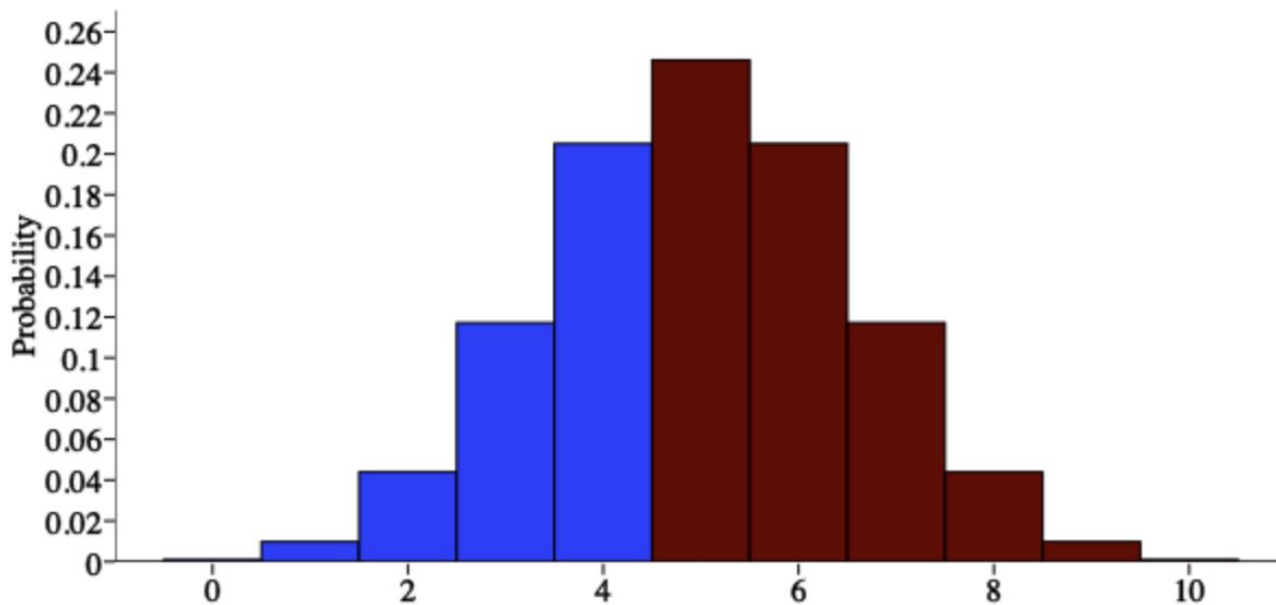
Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.623

n = p =



Confidence check: this probability is > the 0.4512 probability of the 5-success and 6-success red bars, which makes sense

Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

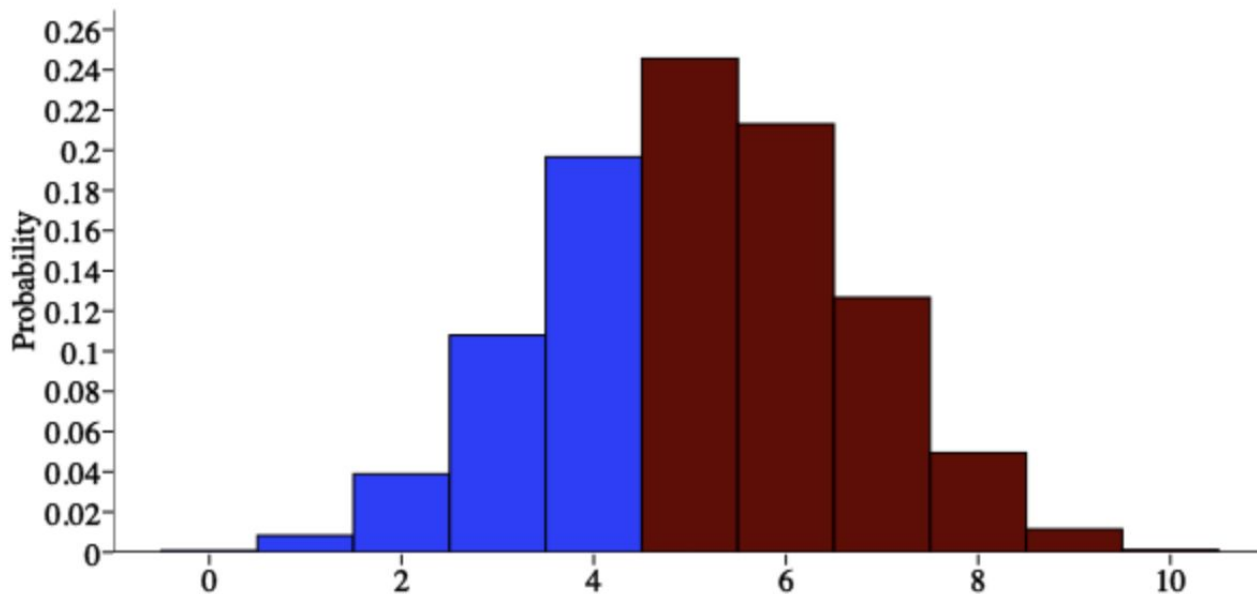
Probability = 0.623

n = 10

p = 0.51

Plot Distribution

What happens if we increase p?



Export this graph

Calculate the probability of successes. successes.

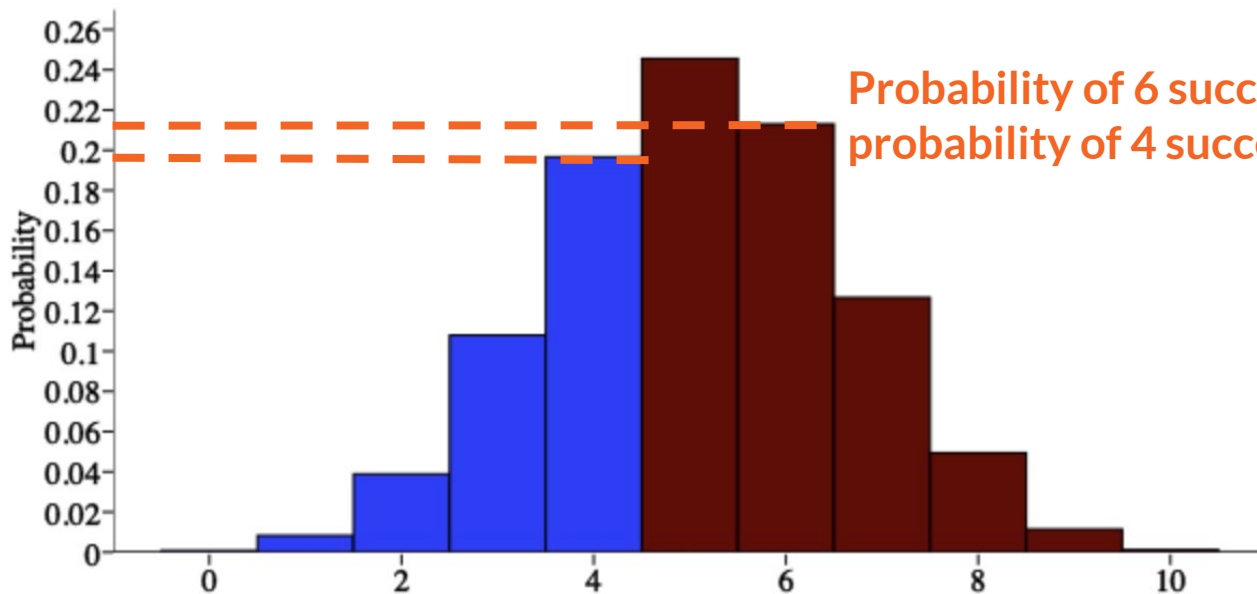
-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.6474

n = 10 p = 0.51 Plot Distribution

What happens if we increase p?



Probability of 6 successes is now > probability of 4 successes

Export this graph

Calculate the probability of at least 5 successes. Go!

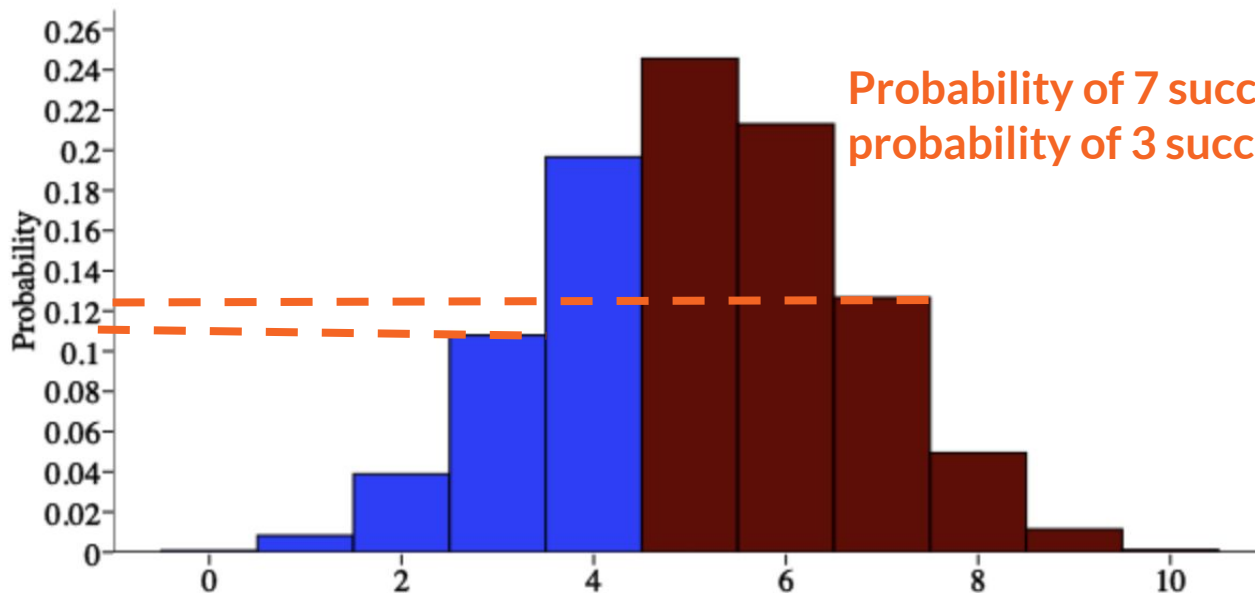
-OR-

Calculate the probability of between and successes (inclusive). Go!

Probability = 0.6474

n = 10 p = 0.51 Plot Distribution

What happens if we increase p?



Probability of 7 successes is now > probability of 3 successes

Export this graph

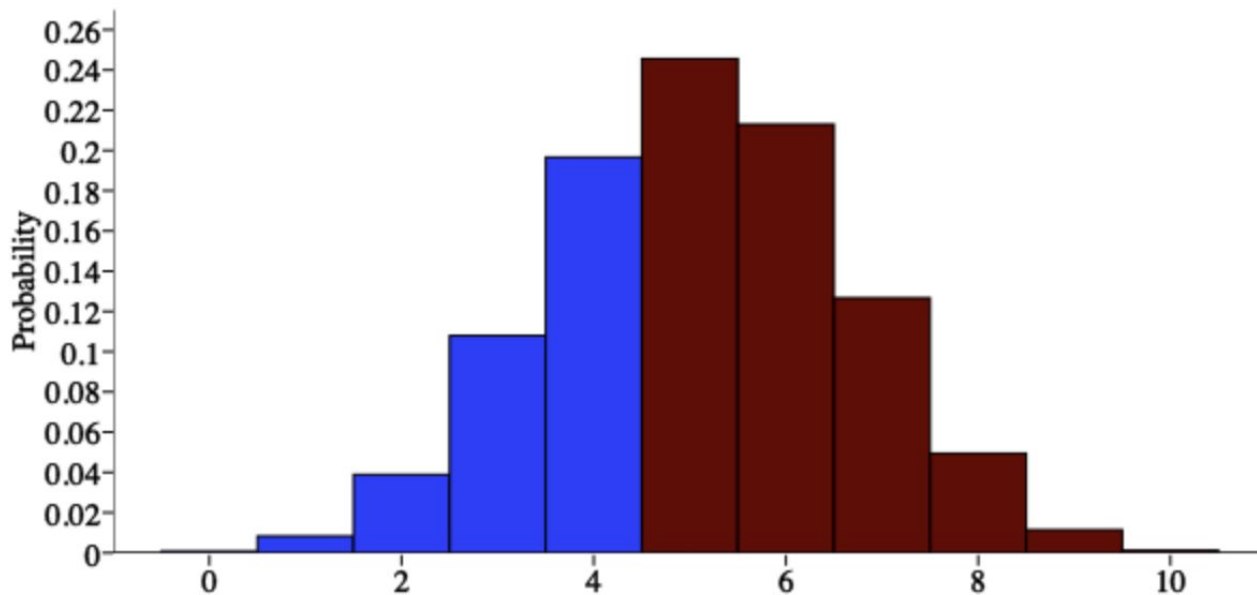
Calculate the probability of at least 5 successes. Go!

-OR-

Calculate the probability of between and successes (inclusive). Go!

Probability = 0.6474

n = p =



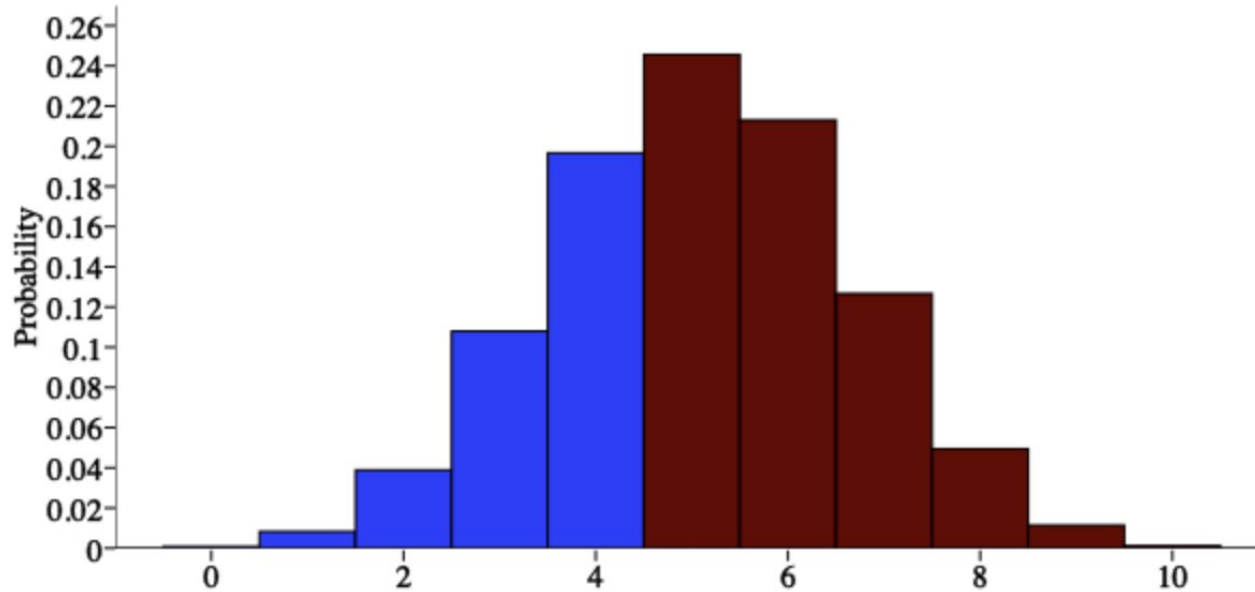
For $p = 0.5$, $P(X \geq 5) = 0.623$. For $p > 0.5$, will $P(X \geq 5)$ be higher or lower than 0.623?

Calculate the probability of successes. successes.

-OR-

Calculate the probability of between and successes (inclusive).

n = p =



For $p = 0.5$, $P(X=5) = 0.623$. For $p > 0.5$, will $P(X=5)$ will be higher.

Calculate the probability of successes.

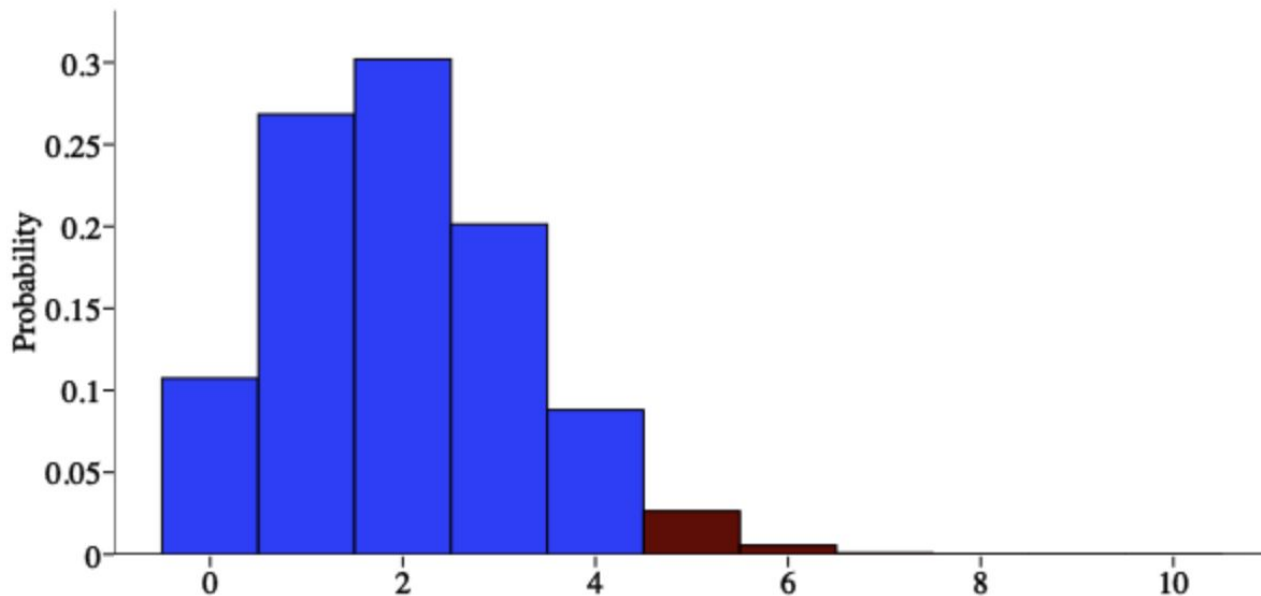
-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.6474

n = p =

What happens if we decrease p?



Calculate the probability of successes.

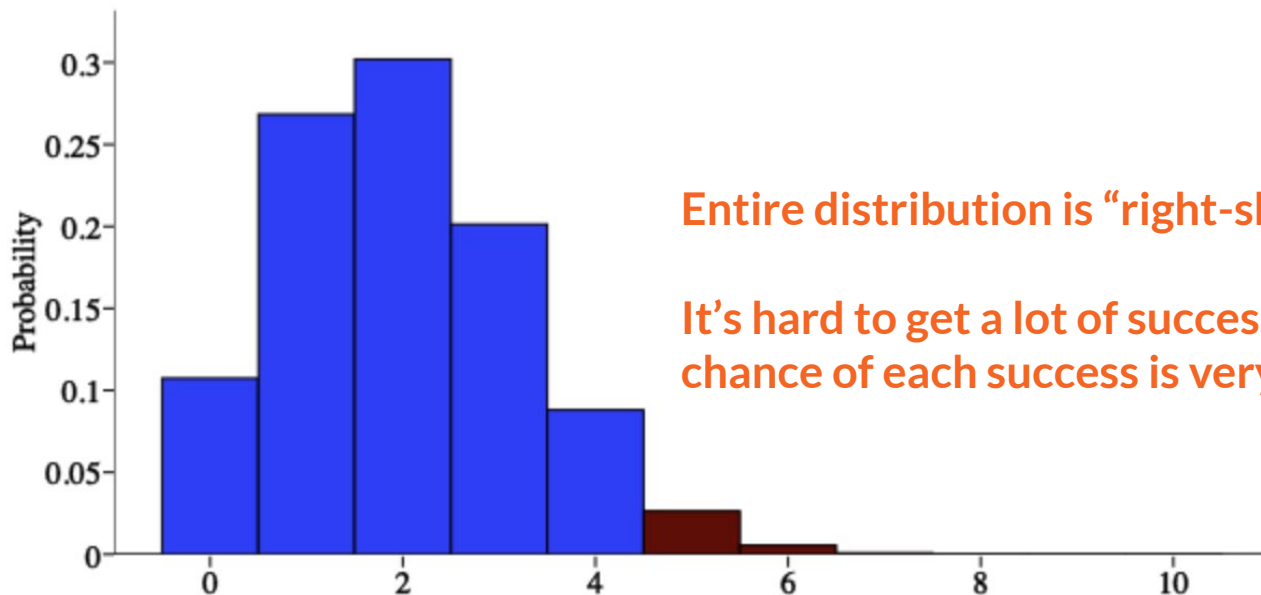
-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.0328

n = p = Prob distribution

What happens if we decrease p?



Entire distribution is “right-skewed”

It's hard to get a lot of successes if the chance of each success is very small!

Export this graph

Calculate the probability of 5 successes.

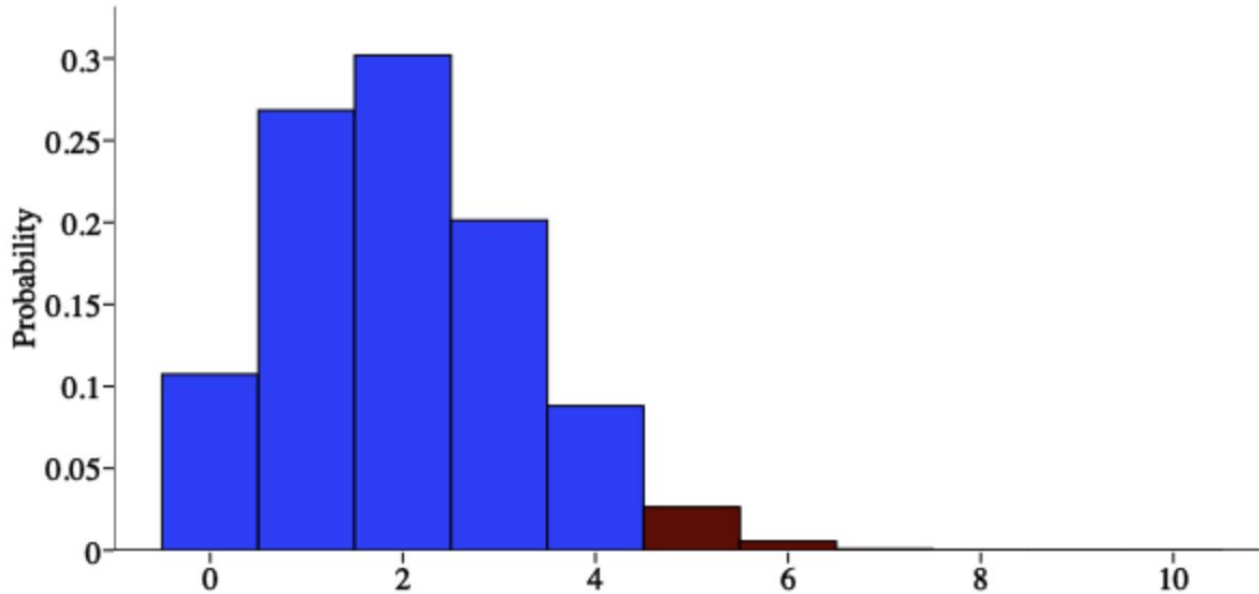
-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.0328

n = 10 p = 0.2 Prob distribution

What happens if we decrease p?



Export this graph

For $p = 0.5$, $P(X \geq 5) = 0.623$. For $p < 0.5$, $P(X \geq 5)$ is lower

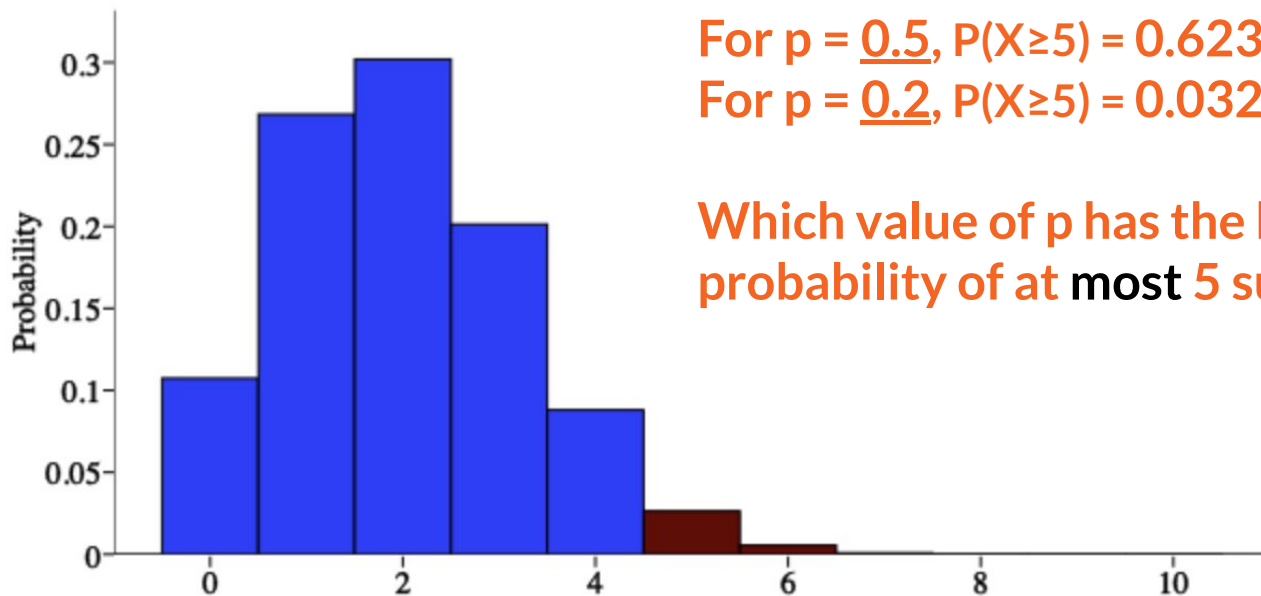
Calculate the probability of at least 5 successes. Go!

-OR-

Calculate the probability of between and successes (inclusive). Go!

Probability = 0.0328

n = p =



For $p = \underline{0.5}$, $P(X \geq 5) = 0.623$.

For $p = \underline{0.2}$, $P(X \geq 5) = 0.0328$.

Which value of p has the higher probability of at **most** 5 successes?

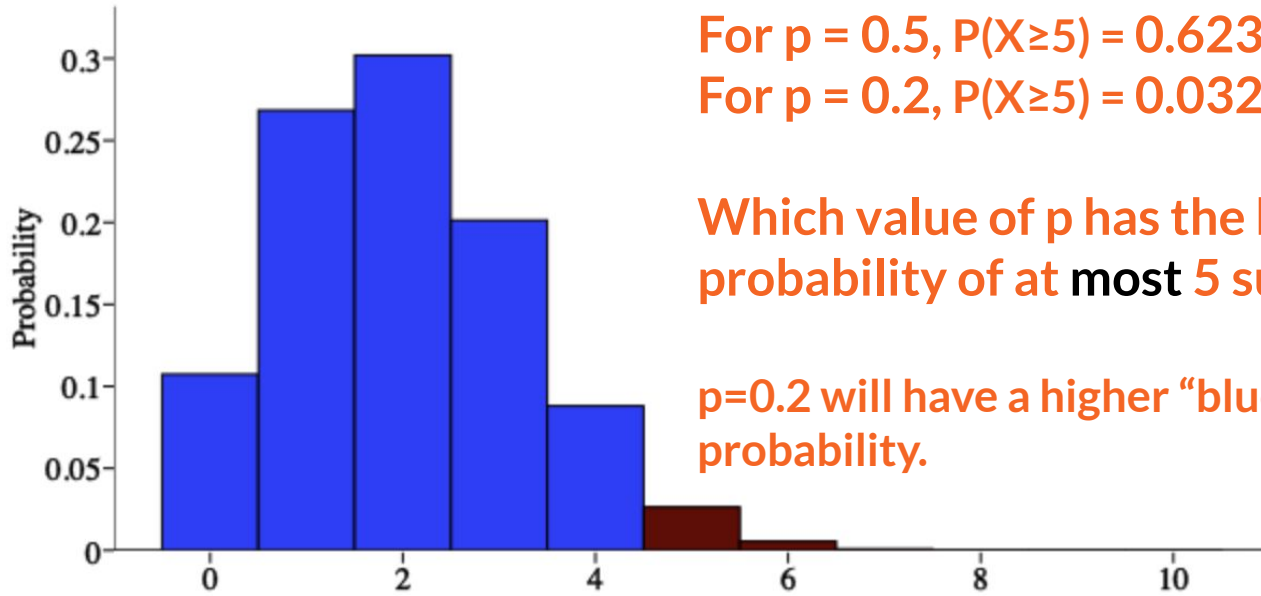
Calculate the probability of successes. successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.0328

n = p =



For $p = 0.5$, $P(X \geq 5) = 0.623$.
For $p = 0.2$, $P(X \geq 5) = 0.0328$.

Which value of p has the higher probability of at **most** 5 successes?

$p=0.2$ will have a higher “blue bar” probability.

Calculate the probability of successes.

-OR-

Calculate the probability of between and successes (inclusive).

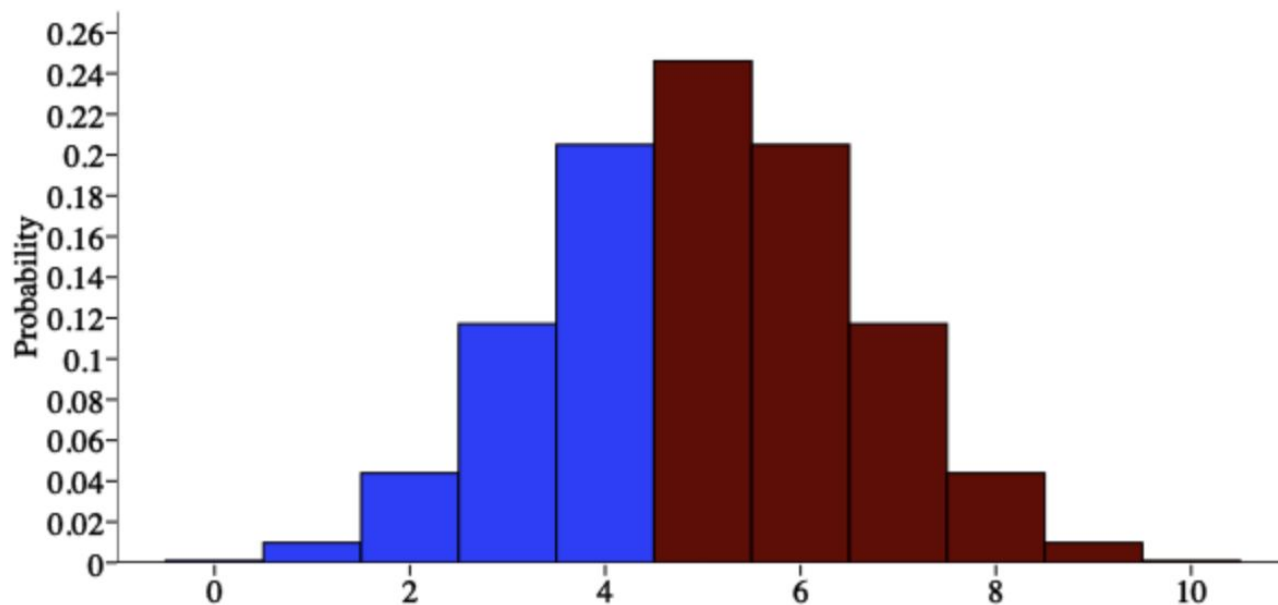
Probability = 0.0328

n = 10

p = 0.5

Prob. distribution

[Back to our original setup](#)



Export this graph

Calculate the probability of 5 successes.

-OR-

Calculate the probability of between and successes (inclusive).

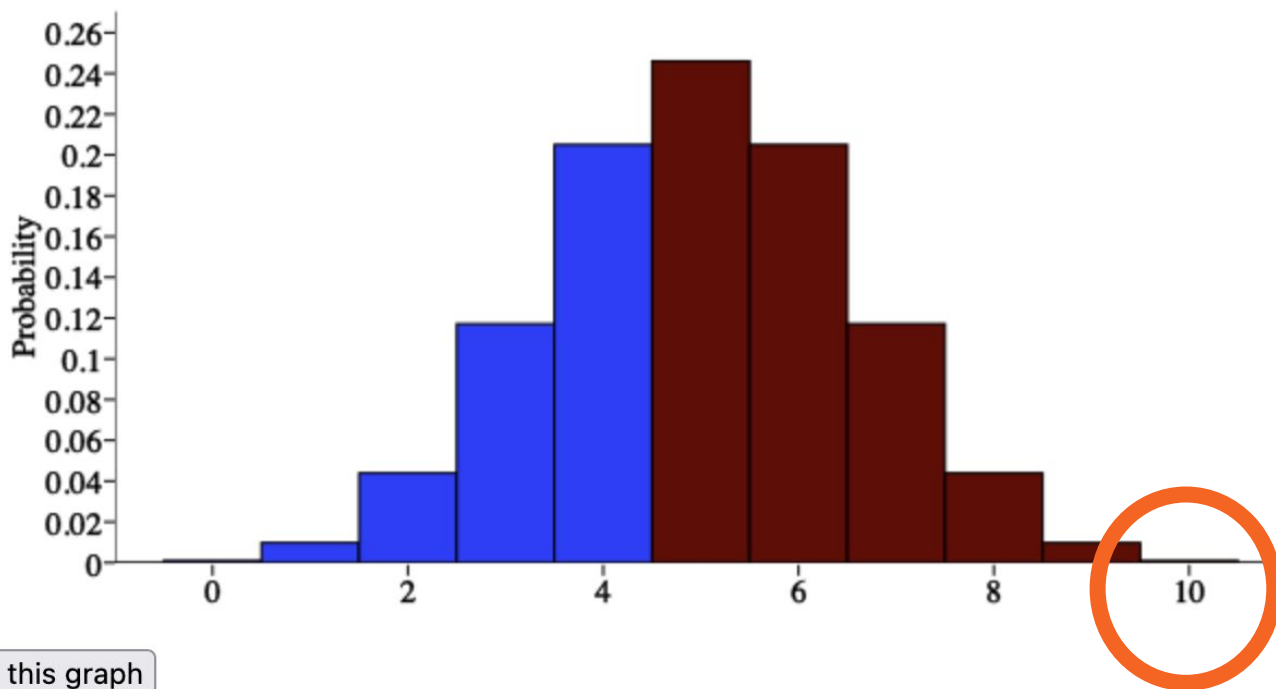
Probability = 0.623

n = 10

p = 0.5

Prob. distribution

Back to our original setup



Calculate the probability of at least 5 successes. Go!

-OR-

Calculate the probability of between and successes (inclusive). Go!

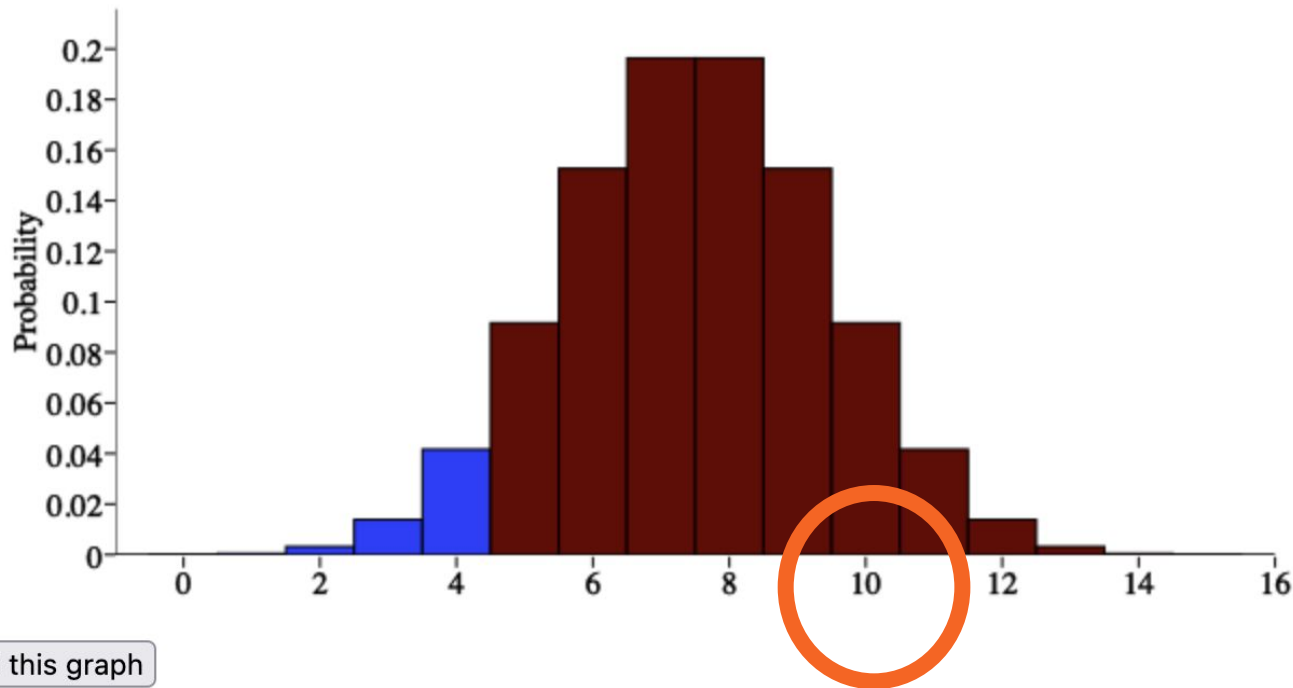
Probability = 0.623

n = 15

p = 0.5

Plot distribution

What happens if we increase n ?



Export this graph

Calculate the probability of at least 5 successes. Go!

-OR-

Calculate the probability of between and successes (inclusive). Go!

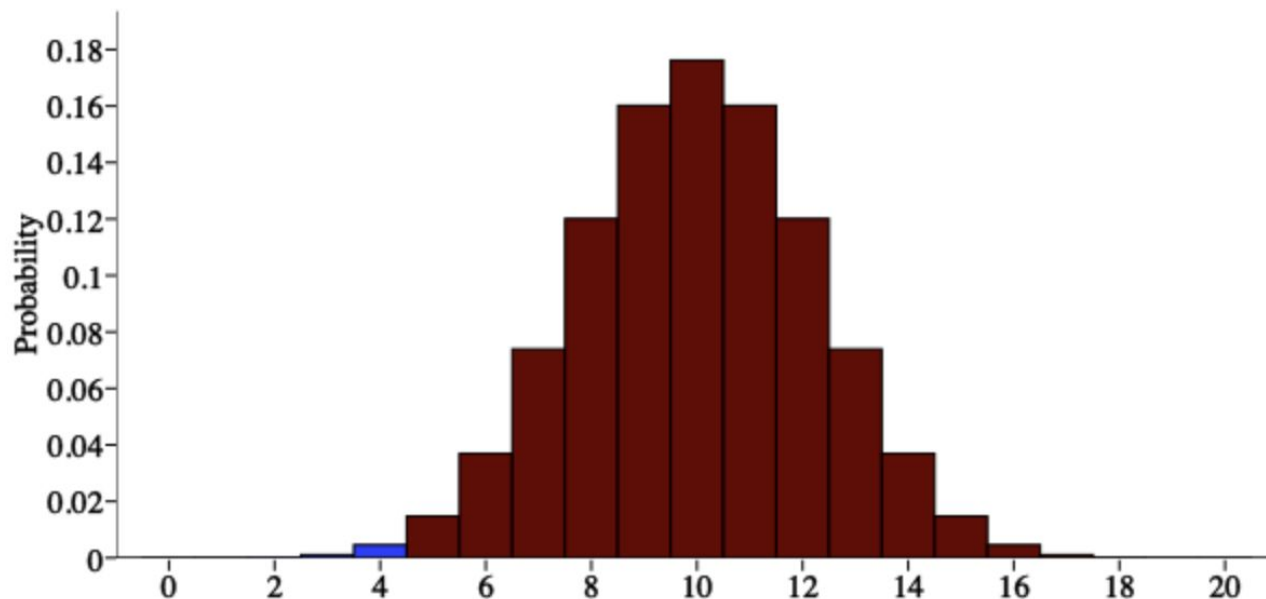
Probability = 0.9408

n = 20

p = 0.5

Plot distribution

What happens if we increase n more?



Export this graph

Calculate the probability of at least 5 successes. Go!

-OR-

Calculate the probability of between and successes (inclusive). Go!

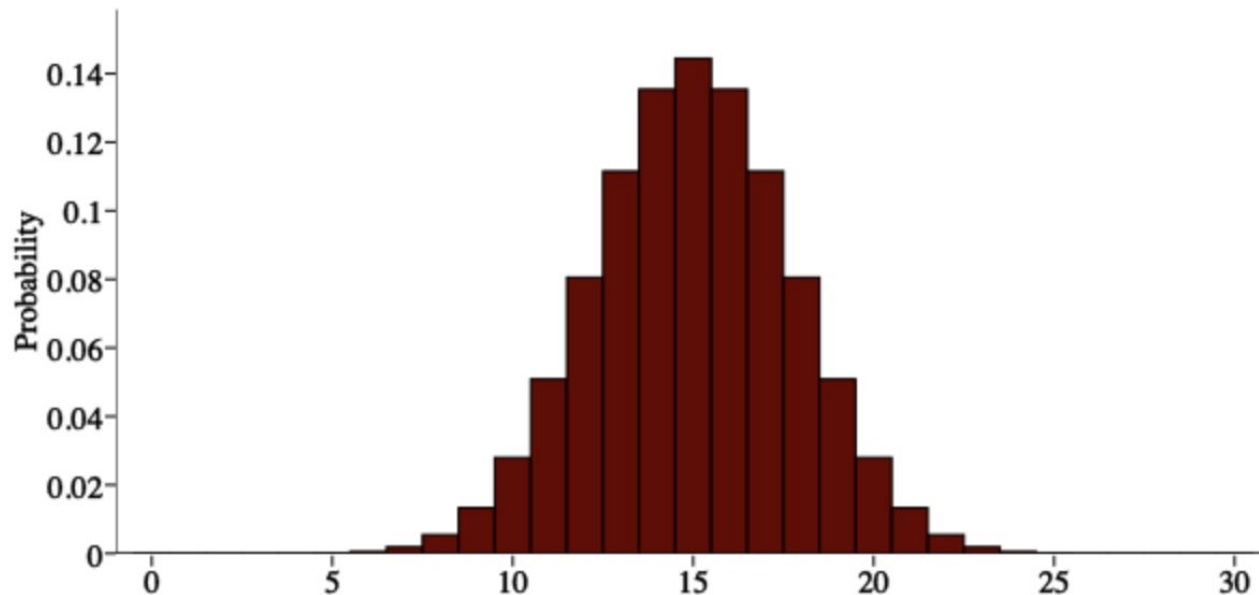
Probability = 0.9941

n = 30

p = 0.5

Plot distribution

What happens if we increase n more?



Export this graph

Calculate the probability of at least 5 successes. Go!

-OR-

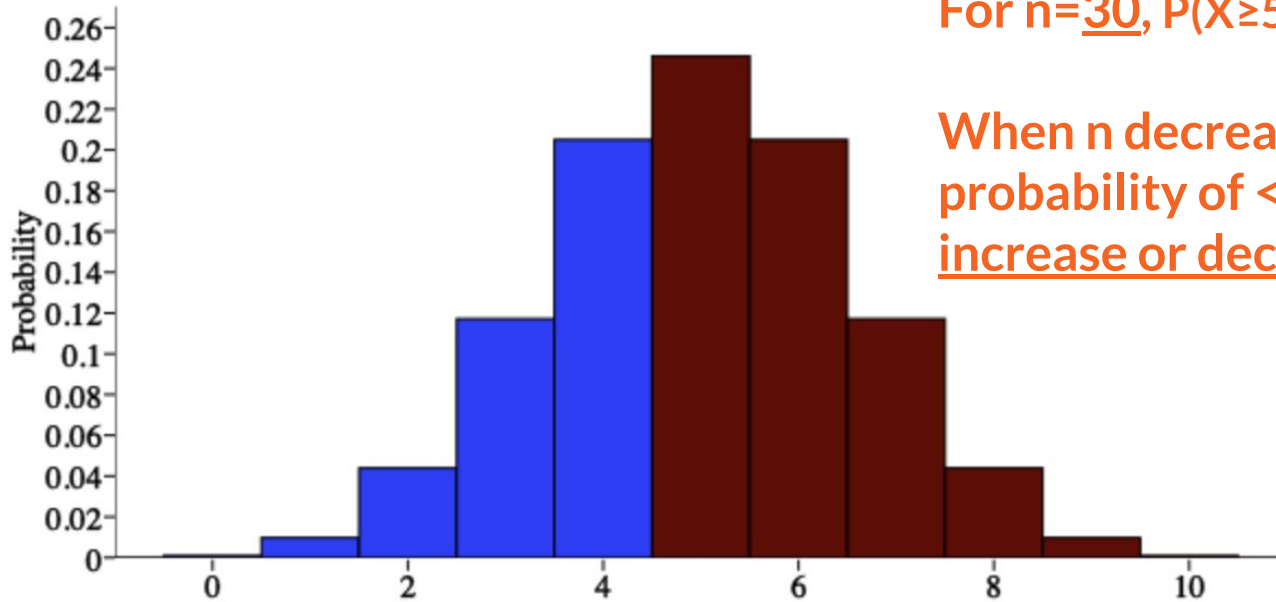
Calculate the probability of between and successes (inclusive). Go!

Probability = 1

n = p =

For n=10, $P(X \geq 5) = 0.623$.

For n=30, $P(X \geq 5) = 1$.



When n decreases, does the probability of < 5 successes increase or decrease?

Calculate the probability of successes. successes.

-OR-

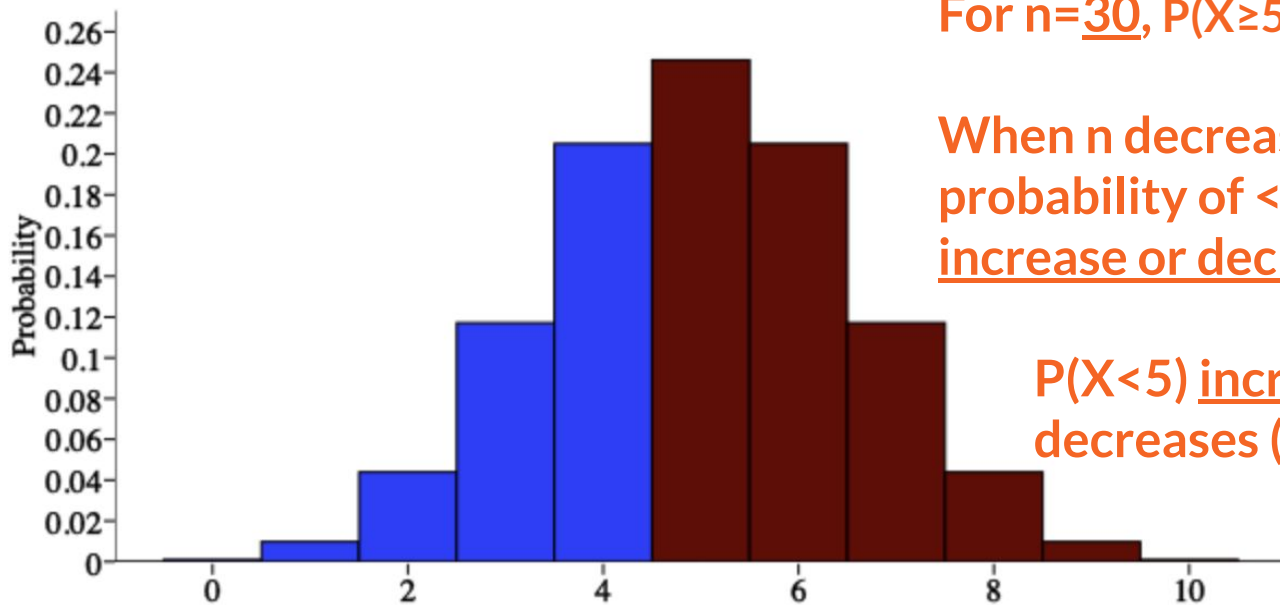
Calculate the probability of between and successes (inclusive).

Probability = 0.623

n = p =

For n=10, $P(X \geq 5) = 0.623$.

For n=30, $P(X \geq 5) = 1$.



When n decreases, does the probability of < 5 successes increase or decrease?

$P(X < 5)$ increases when n decreases (blue bar)

Calculate the probability of successes. successes.

-OR-

Calculate the probability of between and successes (inclusive).

Probability = 0.623

Why does this matter?

- We can talk about the probability with which we think certain sequences will happen
- We can compare different probabilities to each other

Uses of distributions

- They allow us to make precise statements about how uncertain we are
- They help us compare explanations
 - Of two explanations, which is more likely?
 - Does this set of observations need an explanation, or is it just random?
 - What is the single most likely explanation?

1 min break & attendance



tinyurl.com/bp9bm6pd

WHAT WAS THAT NOISE?

IT WAS JUST THE
WIND. OLD
HOUSES MAKE
STRANGE NOISES.
GO BACK TO SLEEP.



IT'S A GHOST! DON'T
LOOK UNDER THE BED!!!

WHAT WAS THAT NOISE?

Null hypothesis

IT WAS JUST THE
WIND. OLD
HOUSES MAKE
STRANGE NOISES.
GO BACK TO SLEEP.



Alternative hypothesis

IT'S A GHOST! DON'T
LOOK UNDER THE BED!!!

WHAT WAS THAT NOISE?

Boring hypothesis

IT WAS JUST THE
WIND. OLD
HOUSES MAKE
STRANGE NOISES.
GO BACK TO SLEEP.



Spooky hypothesis

IT'S A GHOST! DON'T
LOOK UNDER THE BED!!!

Uses of distributions

- Allow us to easily run *hypothesis tests*
 - Use the distribution to tell us which is likelier: the null (boring) hypothesis or the alternative (spooky) hypothesis
 - Can do one-sided or two-sided hypothesis tests, depending on the situation

Tennis refresher

Each tennis player has some number of **serves (S)** per game

A serve may result in several outcomes including an **ace (A)**, where the receiver does not get their racquet on the ball and the server gets a point.

What is the probability that a player will serve an ace in any given serve?

Tennis refresher

Each tennis player has some number of **serves (S)** per game

A serve may result in several outcomes including an **ace (A)**, where the receiver does not get their racquet on the ball and the server gets a point.

What is the probability that a player will serve an ace in any given serve?

Many, many factors, we will ignore (almost) all of them!

Tennis refresher

Each tennis player has some number of **serves (S)** per game

A serve may result in several outcomes including an **ace (A)**, where the receiver does not get their racquet on the ball and the server gets a point.

What is the probability that a player will serve an ace in any given serve?

Factor: being an extremely good tennis player



Competing explanations

1. Serena Williams is equally as good as any other tennis player, someone has to serve the most aces each year
2. Serena Williams is much better than other professional tennis players



Competing explanations

1. Serena Williams is equally as good as any other tennis player, someone has to serve the most aces each year

Women's Tennis Association (WTA) 2012:

~**196,400 Serves (S)** → **11,655 Aces (A)**

Williams 2012 (record # year):

~**2,320 Serves (S)** → **484 Aces (A)**

2. Serena Williams is much better than other professional tennis players



Competing explanations

1. Serena Williams is equally as good as any other tennis player, someone has to serve the most aces each year
2. Serena Williams is much better than other professional tennis players

**Which is null/boring vs.
alternative/spooky?**



Competing explanations

1. Serena Williams is equally as good as any other tennis player, someone has to serve the most aces each year
null/boring
2. Serena Williams is much better than other professional tennis players
alternative/spooky

**Which is null/boring vs.
alternative/spooky?**

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? Answer in terms of A_w , S_w , A_s , or S_s

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? $A_w / S_w = 11,171 / 194,080 = 0.058$

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? $A_w / S_w = 11,171 / 194,080 = 0.058$

What is N ? To figure out whether the average WTA player would have the same number of Aces as Serena, we need to count up to Serena's Serves $S_s = 2,320$

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p? $A_w / S_w = 11,171 / 194,080 = 0.058$

What is N? $S_s = 2,320$

What is the expected number of Aces in N serves?

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? $A_w / S_w = 11,171 / 194,080 = 0.058$

What is N ? $S_s = 2,320$

What is the expected number of Aces in N serves?
(express in terms of p and N)

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? $A_w / S_w = 11,171 / 194,080 = 0.058$

What is N ? $S_s = 2,320$

Expected # Aces in N serves? $N * p = 134.56$

Standard deviation? $\sqrt{N * p * (1-p)} = 11.26$

Binomial distribution statistics

$$\mathbb{E}[X] = N p$$

$$Var[X] = N p(1 - p)$$

$$Std[X] = \sqrt{N p(1 - p)}$$

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? 0.058

What is N ? 2,320

Expected # Aces in N serves? 134.56

Standard deviation? 11.26

Represent *other* WTA players as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

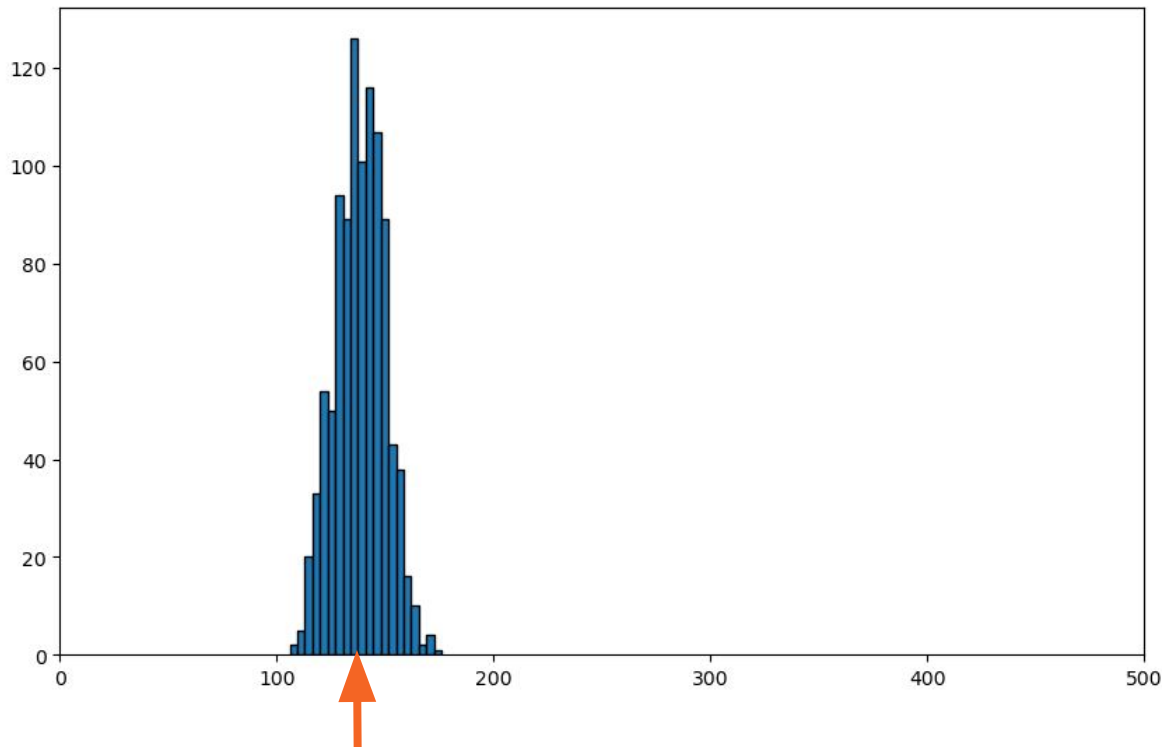
What is p ? 0.058

What is N ? 2,320

Expected # Aces in N serves? 134.56

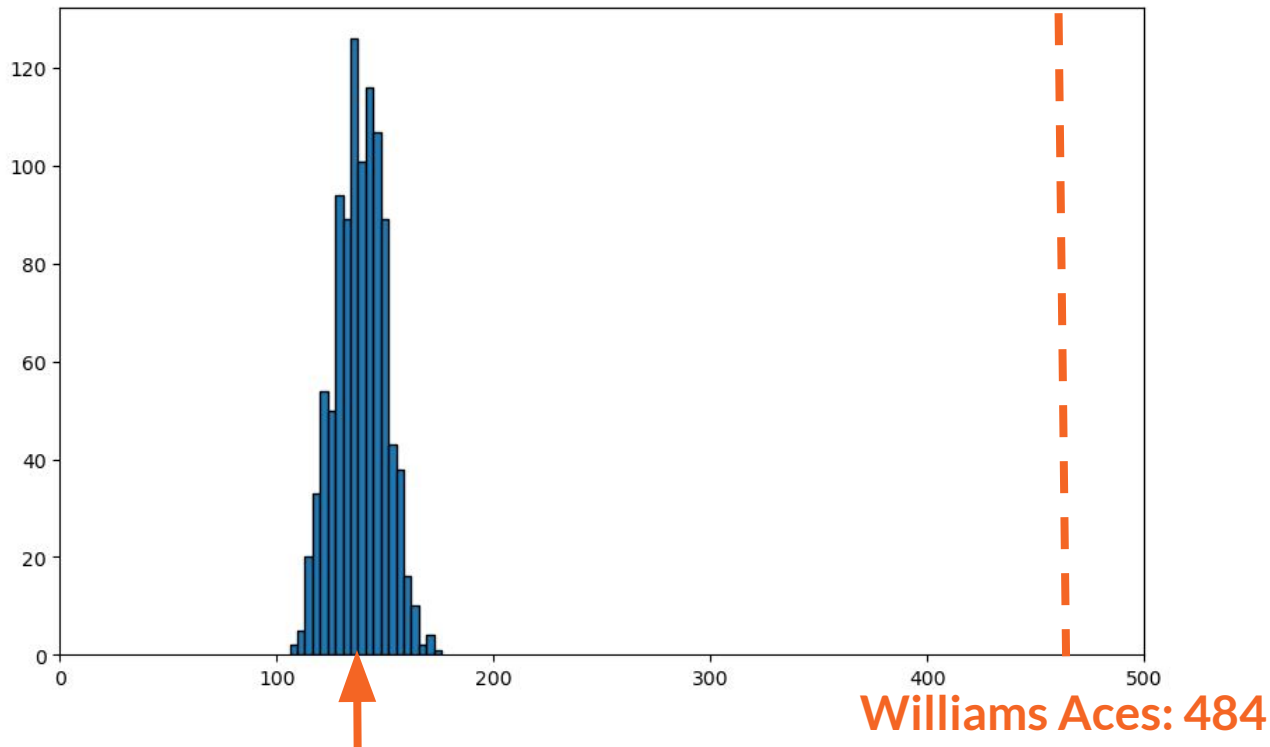
Standard deviation? 11.26

Aces distribution of other WTA players



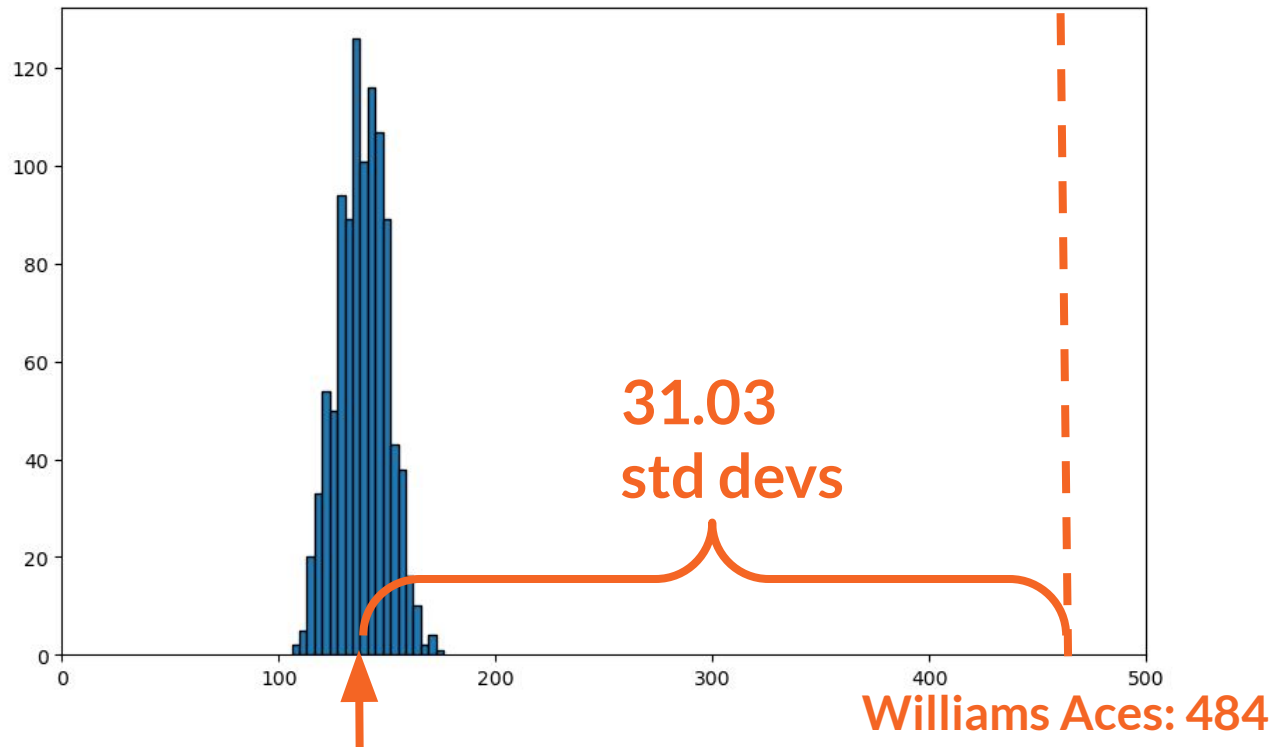
Expected # Aces in N serves **134.56**

Aces distribution of other WTA players



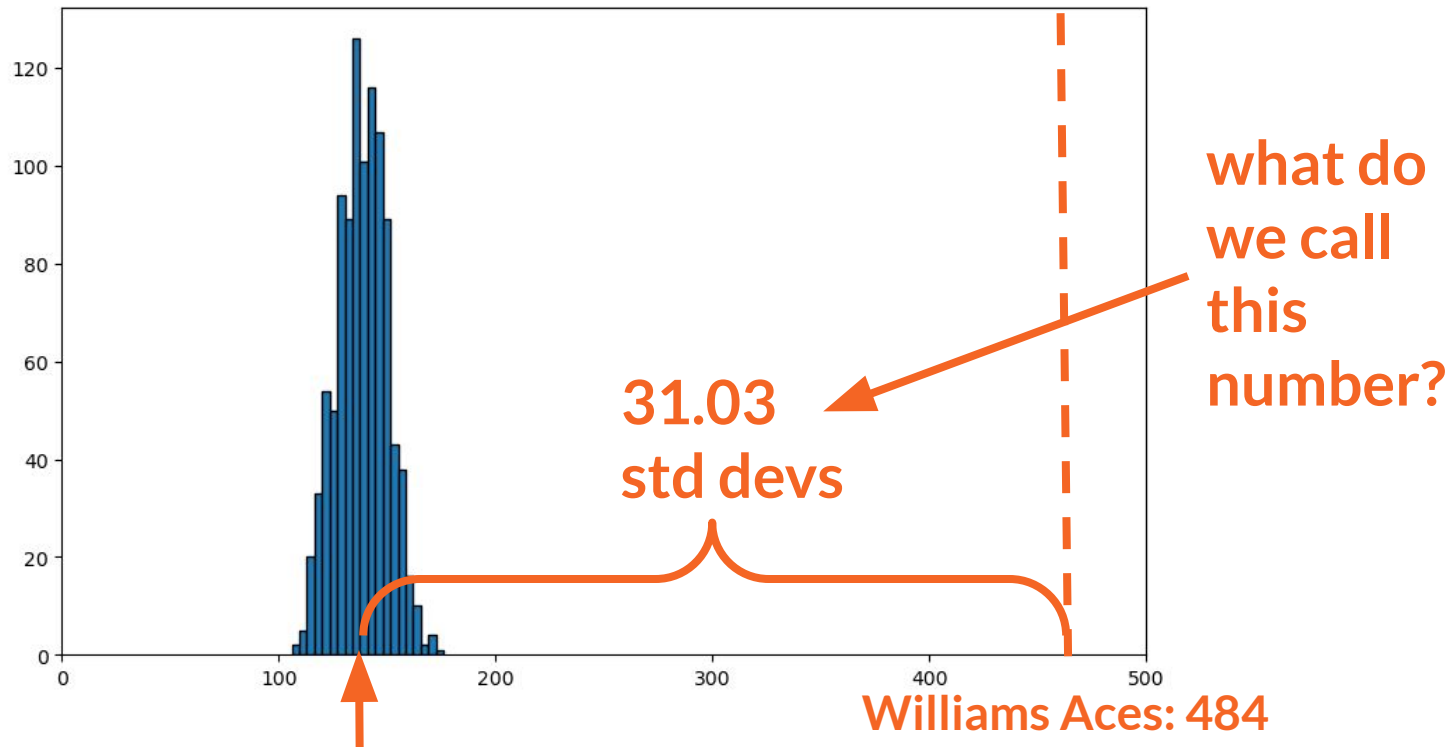
Expected # Aces in N serves 134.56

Aces distribution of other WTA players



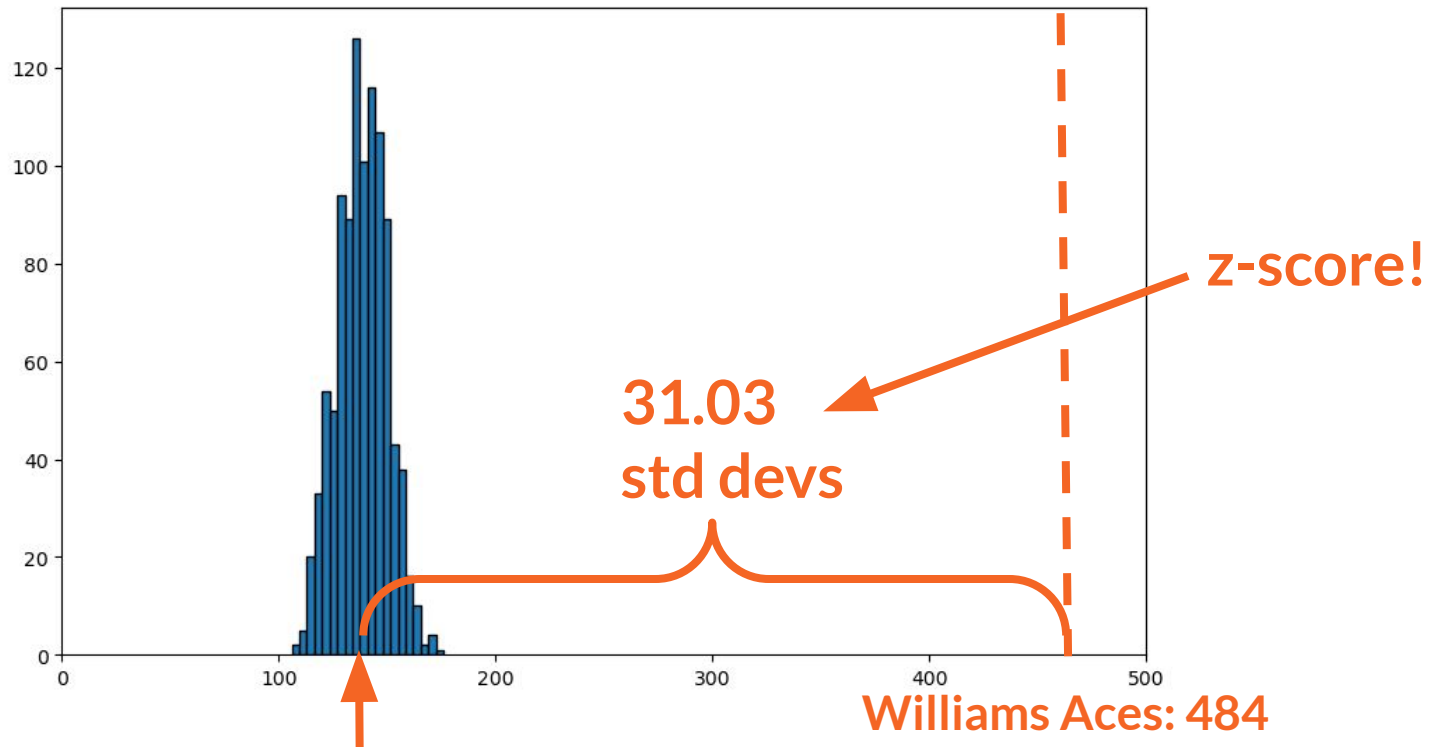
Expected # Aces in N serves 134.56

Aces distribution of other WTA players



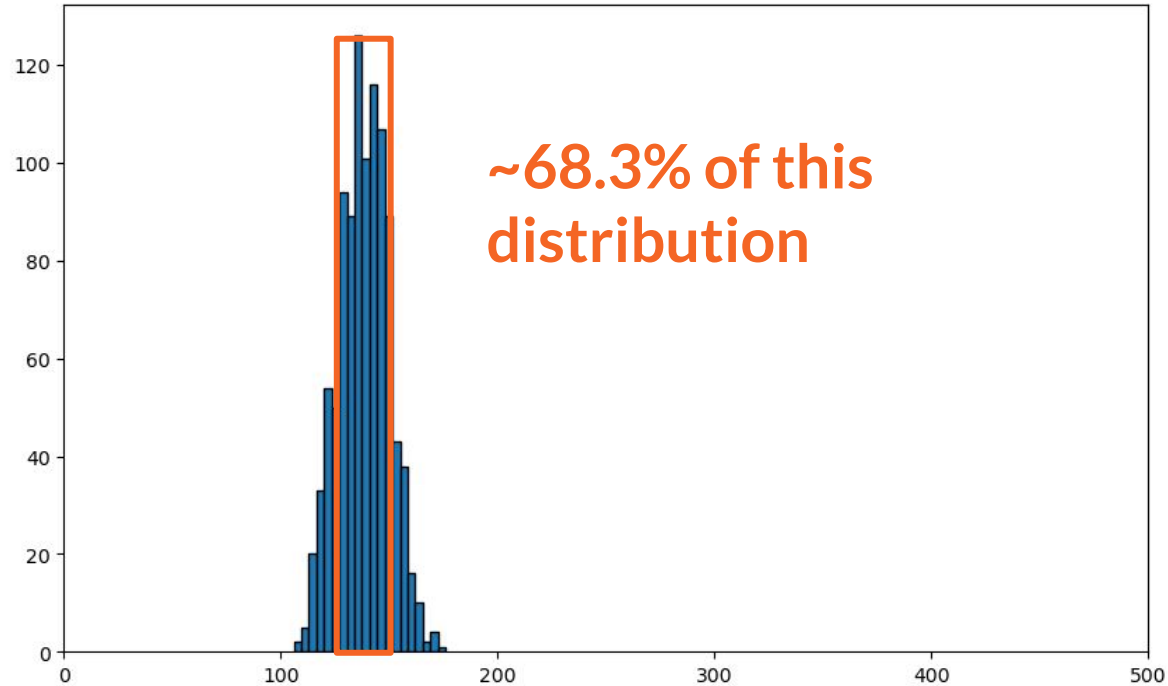
Expected # Aces in N serves 134.56

Aces distribution of other WTA players

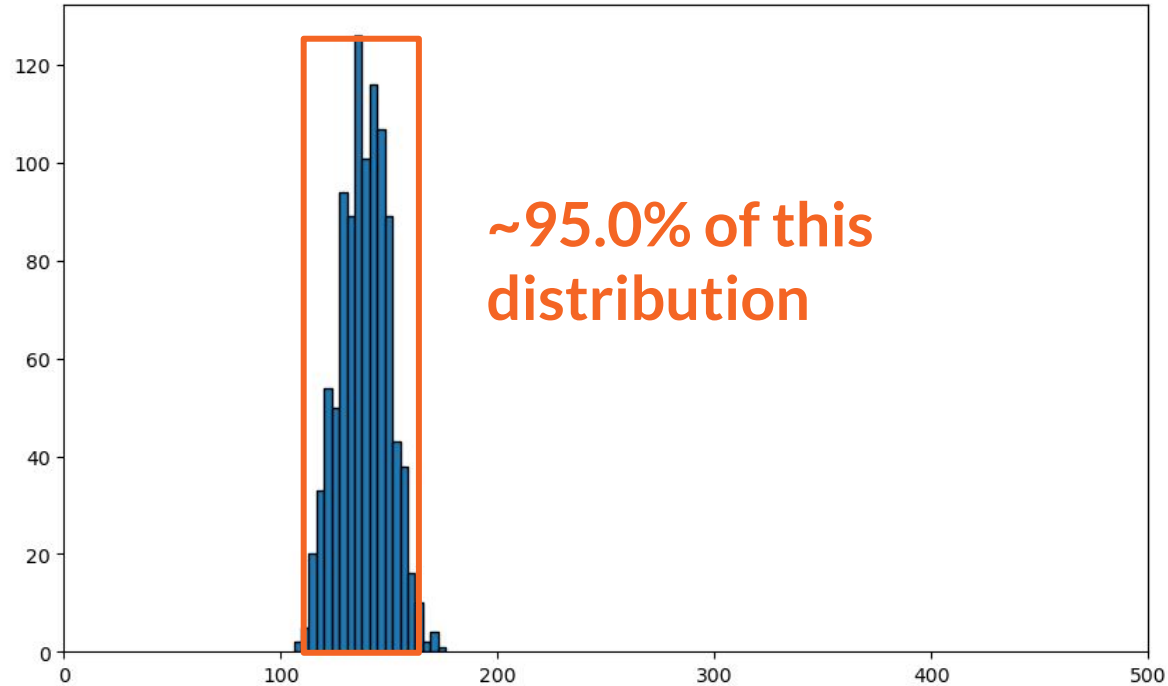


Expected # Aces in N serves 134.56

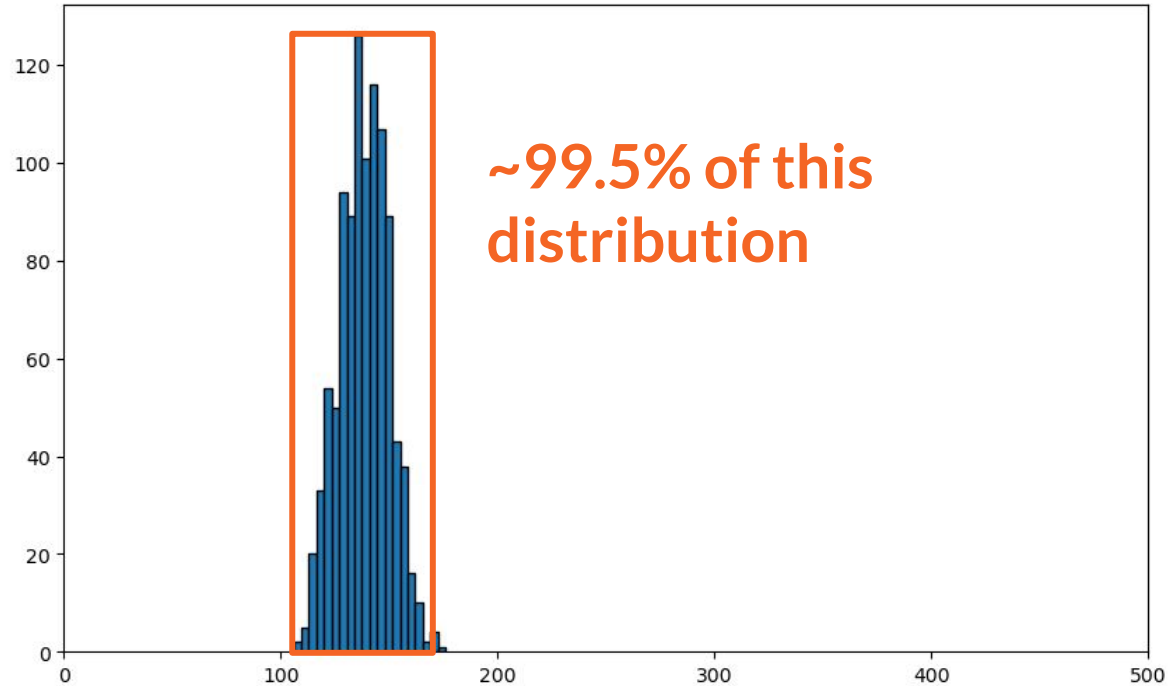
One standard deviation



Two standard deviations



Three standard deviations



The 68/95/99.7 rule

If a distribution is approximately normal:

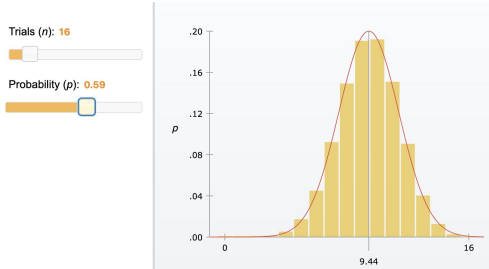
- 68% within ONE standard deviation
- 95% within TWO standard deviations
- 99.7% within THREE standard deviations
- Almost nothing outside 3 sd

The 68/95/99.7 rule

If a distribution is approximately normal:

- 68% within ONE standard deviation
- 95% within TWO standard deviations
- 99.7% within THREE standard deviations
- Almost nothing outside 3 sd

When N is large and p is not close to 0 or 1, binomial is approximately normal



Normals in numpy



Molly White
@molly0xFF

2:07 PM · Oct 7, 2023 · 2.1M Views ...

From yesterday's exhibits in US v. Sam Bankman-Fried:

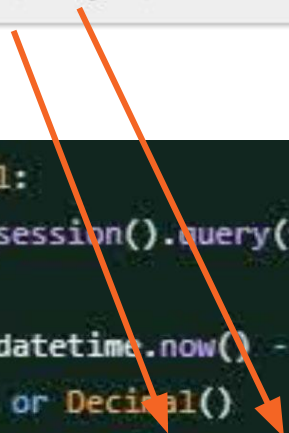
The prosecution shows that the "insurance fund" that FTX bragged about was fake, and just calculated by multiplying daily trading volume by a random number around 7500

```
def _get_change() -> Decimal:
    daily_volume = current_session().query(func.sum(Trade.size
* Trade.price)).filter(
        Trade.created_at > datetime.now() -
timedelta(days=1)).scalar() or Decimal()
    return f2d(numpy.random.normal(7500, 3000)) * daily_volume
/ Decimal('1e9')
```

Normals in numpy

```
>>> mu, sigma = 0, 0.1 # mean and standard deviation
>>> s = np.random.normal(mu, sigma, 1000)
```

```
def _get_change() -> Decimal:
    daily_volume = current_session().query(func.sum(Trade.size
* Trade.price)).filter(
        Trade.created_at > datetime.now() -
timedelta(days=1)).scalar() or Decimal()
    return f2d(numpy.random.normal(7500, 3000)) * daily_volume
/ Decimal('1e9')
```



Normals in numpy

Did they import numpy as np?

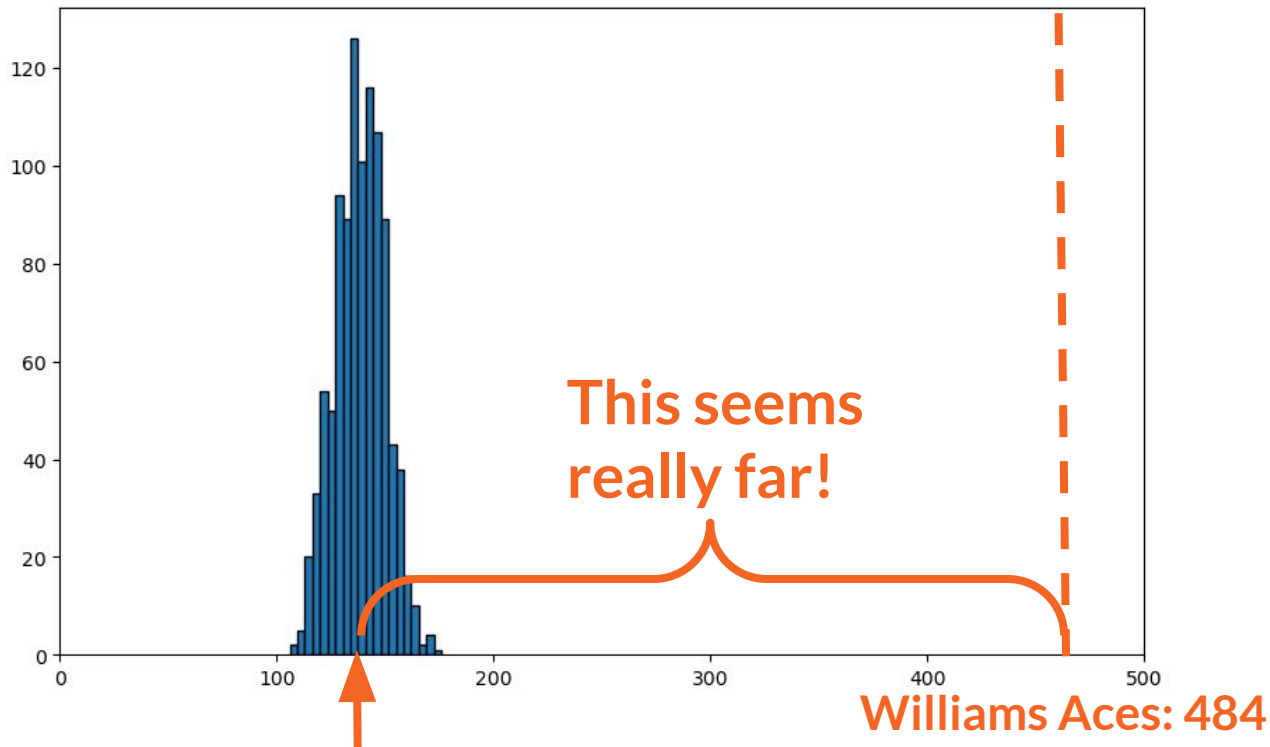
```
def _get_change() -> Decimal:
    daily_volume = current_session().query(func.sum(Trade.size
* Trade.price)).filter(
        Trade.created_at > datetime.now() -
timedelta(days=1)).scalar() or Decimal()
    return f2d(numpy.random.normal(7500, 3000)) * daily_volume
/ Decimal('1e9')
```

Normals in numpy

Did they import numpy as np? **No! Otherwise this would be `np.random.normal()`**

```
def _get_change() -> Decimal:
    daily_volume = current_session().query(func.sum(Trade.size
* Trade.price)).filter(
        Trade.created_at > datetime.now() -
timedelta(days=1)).scalar() or Decimal()
    return f2d(numpy.random.normal(7500, 3000)) * daily_volume
/ Decimal('1e9')
```


Aces distribution of other WTA players



Expected # Aces in N serves 134.56

Represent Serena Williams as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ?

What is N ?

Answer in terms of A_s or S_s

Expected # Aces in N serves?

Standard deviation?

Represent Serena Williams as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ?

What is N ? $S_s = 2,320$

Expected # Aces in N serves? $A_s = 484$

Standard deviation?

Represent Serena Williams as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? $A_s/N = 484 / 2320 = 0.209$

What is N ? $S_s = 2,320$

Expected # Aces in N serves? $A_s = 484$

Standard deviation?

Represent Serena Williams as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

What is p ? $A_s/N = 484 / 2320 = 0.209$

What is N ? $S_s = 2,320$

Expected # Aces in N serves? $A_s = 484$

Standard deviation? $\text{sqrt}(N * p * (1-p)) = 19.57$

Represent Serena Williams as binomial

Other WTA players: 11,171 aces (A_w) in 194,080 serves (S_w)

Serena Williams: 484 Aces (A_s) in 2,320 Serves (S_s)

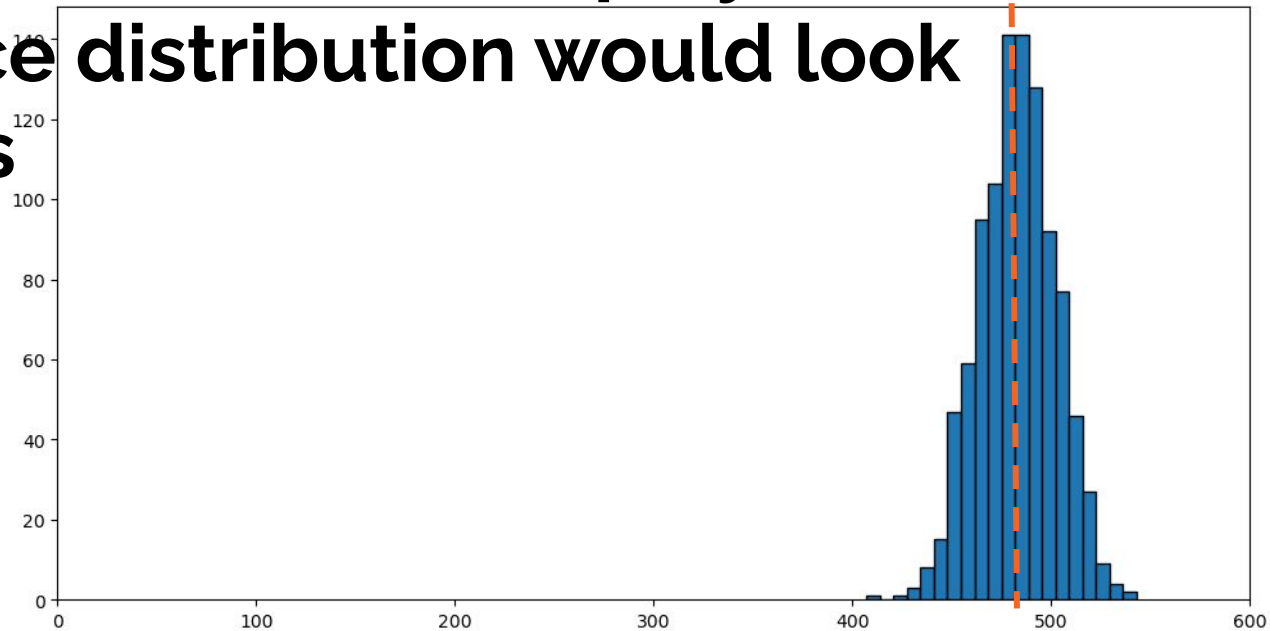
What is p ? 0.209

What is N ? 2,320

Expected # Aces in N serves? 484

Standard deviation? 19.57

If William's probability $p=0.209$ is representative of WTA players, their ace distribution would look like this



Serena Williams

What can we say about Williams?

Is her true Aces/Serves probability = 20.9%?

Is her true Aces/Serves probability = 5.8%?

What can we say about Williams?

Is her true Aces/Serves probability = 20.9%?

Maybe? Could be a little high

Is her true Aces/Serves probability = 5.8%?

What can we say about Williams?

Is her true Aces/Serves probability = 20.9%?

Maybe? Could be a little high

Is her true Aces/Serves probability = 5.8%?

Definitely not. No chance.

What can we say about Williams?

We don't know
enough to
accept the
alternative
hypothesis

Is her true Aces/Serves probability = 20.9%?

Maybe? Could be a little high

Is her true Aces/Serves probability = 5.8%?

What can we say about Williams?

Is her true Aces/Serves probability = 20.9%?

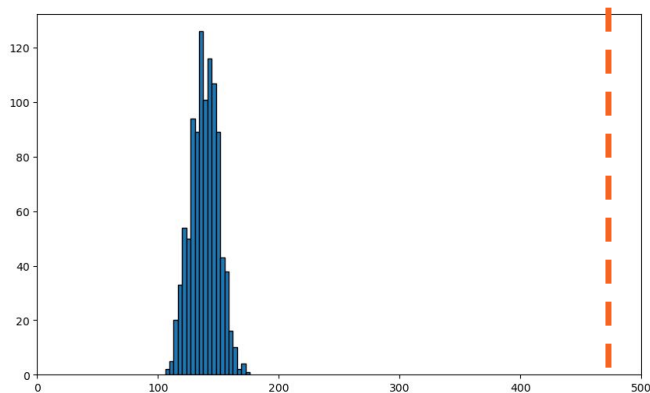
Maybe? Could be a little high

Is her true Aces/Serves probability = 5.8%?

But we can
reject the null
hypothesis

Definitely not. No chance.

What do we need for a hypothesis test?

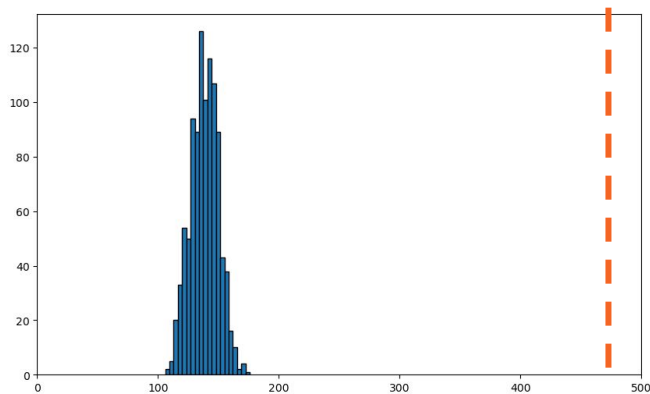


- Null (boring) hypothesis H_0
- Alternative (spooky) hypothesis H_1

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Serena
Williams

What do we need for a hypothesis test?

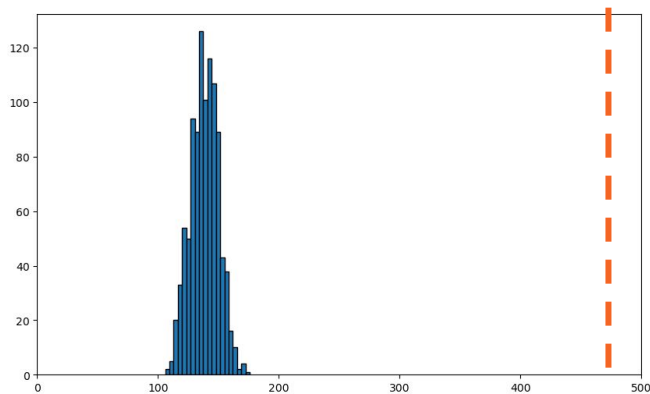


$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Serena
Williams

- Null (boring) hypothesis H_0
- Alternative (spooky) hypothesis H_1
- A hypothesis test gets us from an observed number to some sort of standardized “spooky score”

What do we need for a hypothesis test?

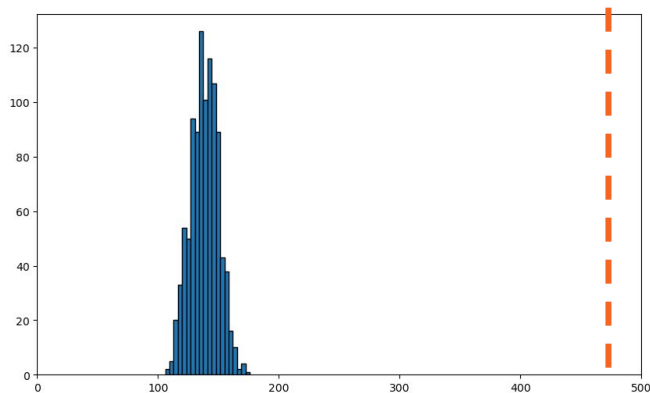


$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Serena
Williams

- Null (boring) hypothesis H_0
- Alternative (spooky) hypothesis H_1
- Plot where your value is relative to the distribution
- Decide whether you **reject the null**
(“Tradition” says: at alpha level 5%, do you have a z-score greater than [the z-score reference from z-table, 1.645]? If yes, reject the null – it’s spooky)

What do we need for a hypothesis test?



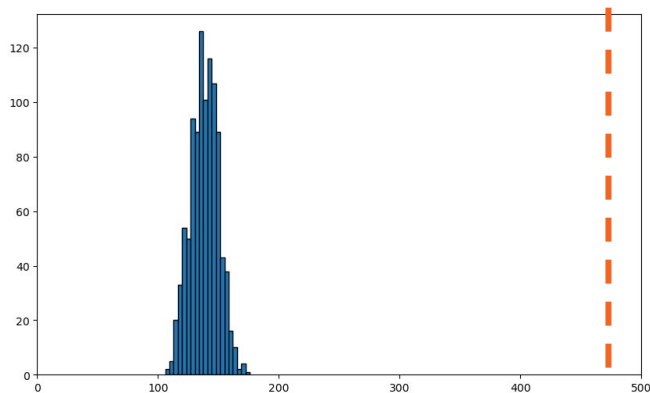
$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Serena
Williams

Alpha: the probability of rejecting the null hypothesis when it was in fact true

(“Tradition” says: at **alpha** level 5%, do you have a z-score greater than [the z-score reference from z-table, 1.645]? If yes, reject the null – it’s spooky)

What do we need for a hypothesis test?



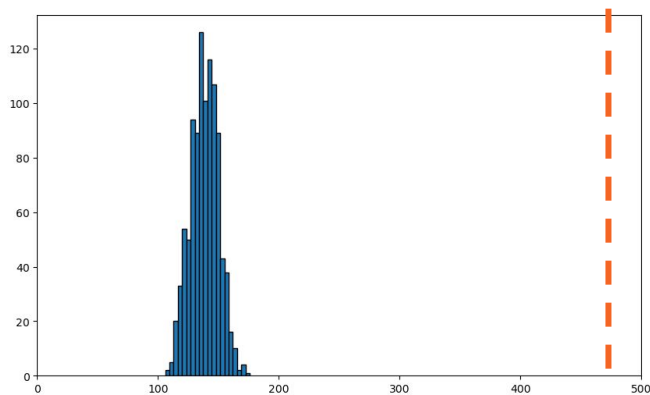
$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Serena
Williams

Alpha: the probability of rejecting the null hypothesis when it was in fact true (also known as probability of Type I Error or **False Positive**!)

(“Tradition” says: at **alpha** level 5%, do you have a z-score greater than [*the z-score reference from z-table, 1.645*]? If yes, reject the null – it’s spooky)

What do we need for a hypothesis test?



$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

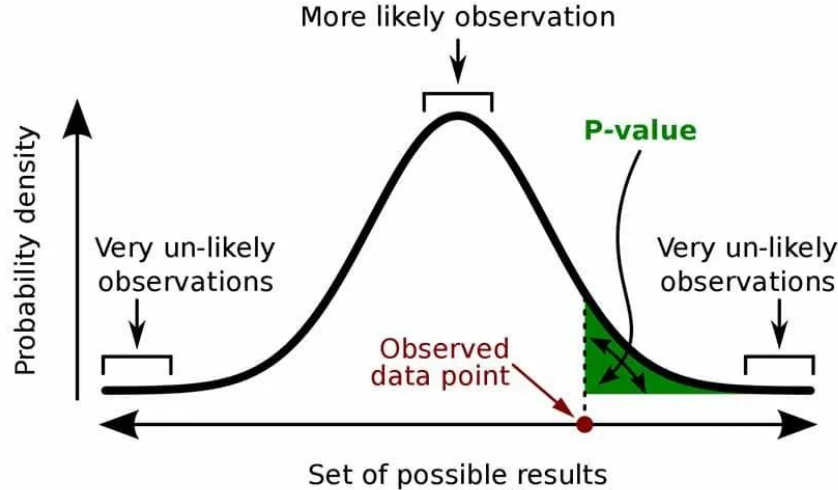
Serena
Williams

If your p-value is greater than alpha, you cannot reject the null hypothesis. (*Boring case!*)

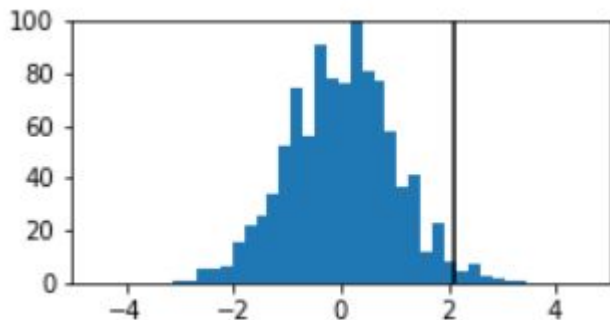
If your p-value is less than alpha, you can reject the null hypothesis. (*Spooky case!*)

What is a p-value?

- The probability that something occurred, given the null hypothesis is true
 - Lower p-value → spookier → more “statistical significance” of your observation
 - The cumulative distribution function (under the null hypothesis) at the observed x-value



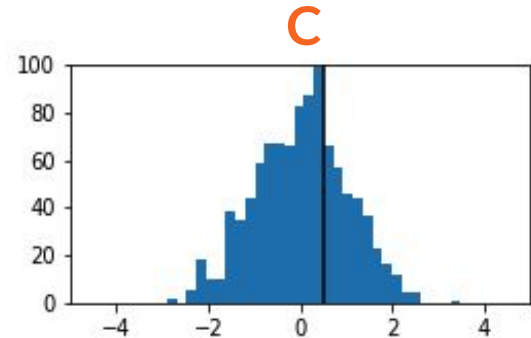
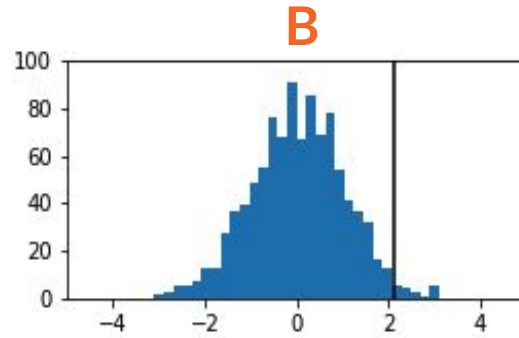
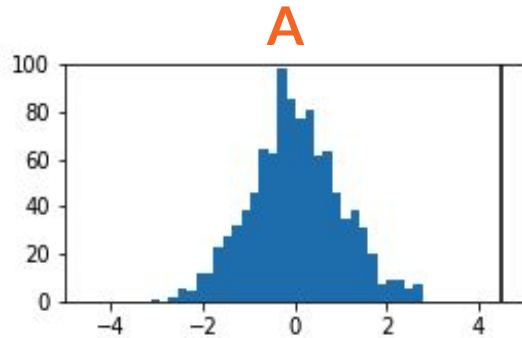
Is 5% a magic number?



$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

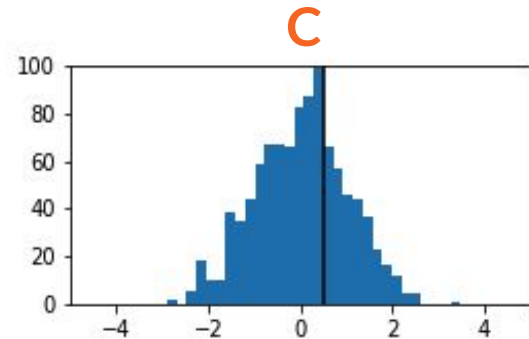
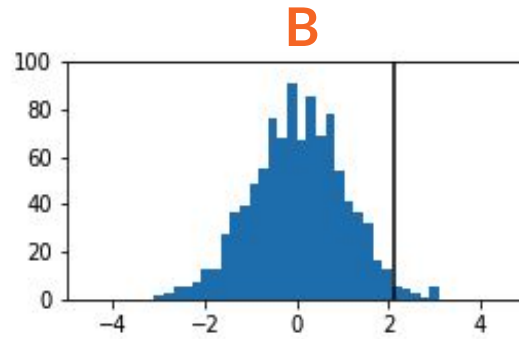
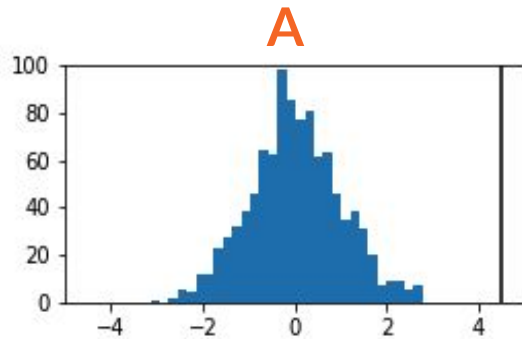
- Would someone at the black vertical line here ($p=0.04$) necessarily be considered “spooky”?
 - Not necessarily!
- 5% is a useful benchmark but not a hard and fast rule

Match the figures to p-values



p-values: 0.04, 0.000001, 0.6

Match the figures to p-values



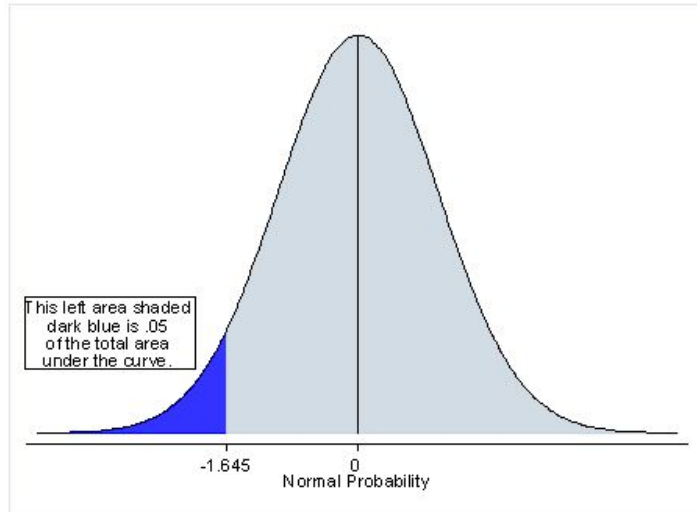
p-values: 0.04, 0.000001, 0.6

B

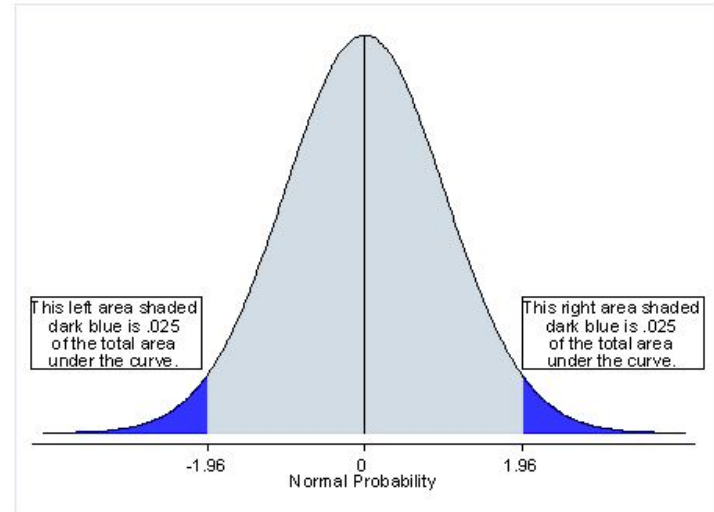
A

C

One-tailed vs. Two-tailed

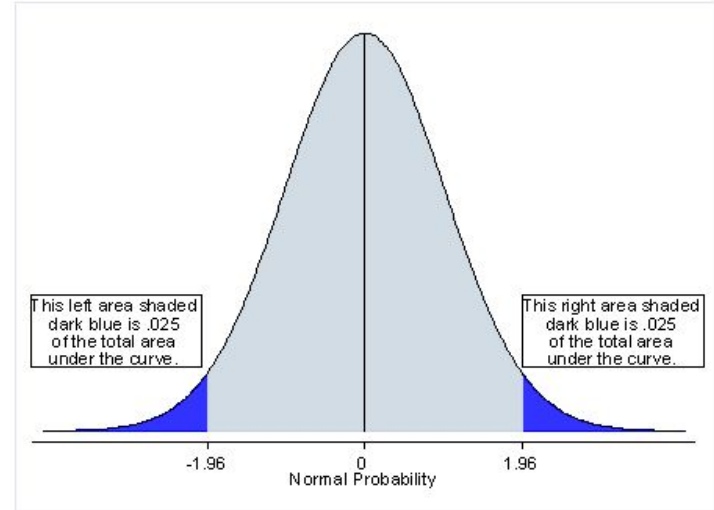
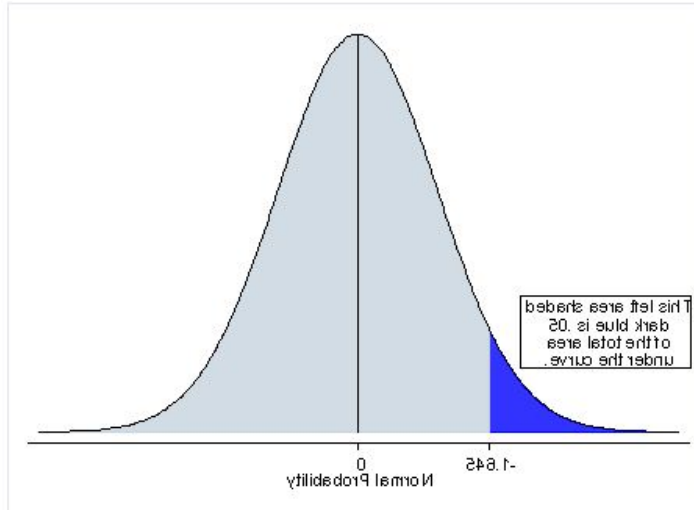


One-tailed: if something is either *better/worse*



Two-tailed: if something is *different (in either direction)*

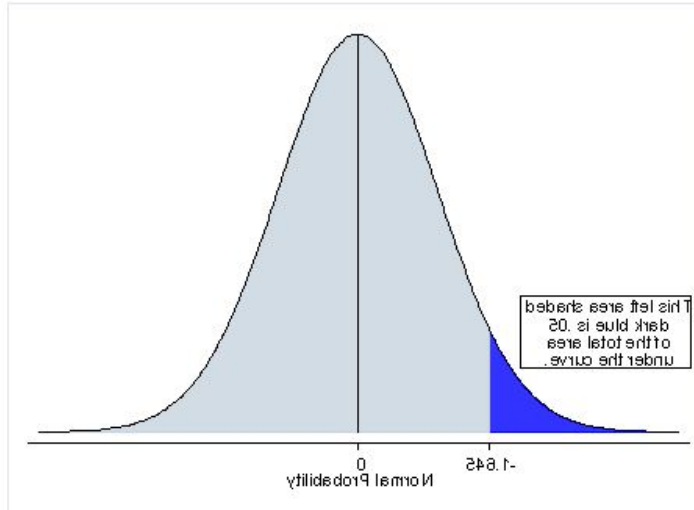
Null (H_0) vs Alternative (H_a) Hypotheses



H_0 : Serena Williams is not significantly better than the average WTA player: $A_S / S_S \leq A_W / S_W$

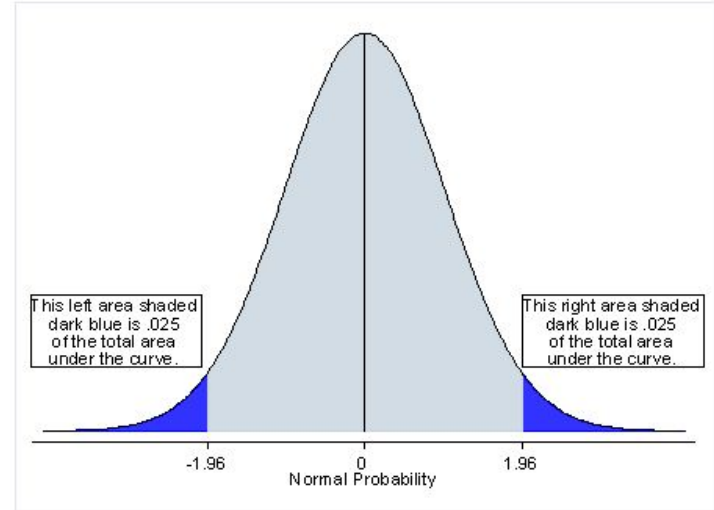
H_a : Serena Williams is significantly better than the average WTA player: $A_S / S_S > A_W / S_W$

Null (H_0) vs Alternative (H_a) Hypotheses



H_0 : Serena Williams is not significantly better than the average WTA player: $A_S / S_S \leq A_W / S_W$

H_a : Serena Williams is significantly better than the average WTA player: $A_S / S_S > A_W / S_W$

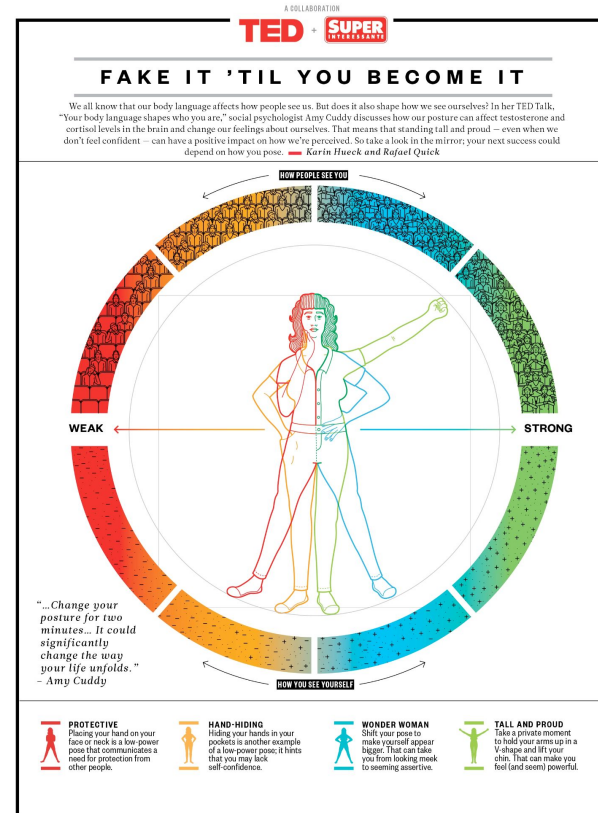


H_0 : there is no significant difference between Serena Williams and the average WTA player ($A_S / S_S = A_W / S_W$)

H_a : $A_S / S_S \neq A_W / S_W$



Power Posing



Power Posing: P-hacked

Power Posing: P-Curving the Evidence

Joseph P. Simmons and Uri Simonsohn

University of Pennsylvania

Abstract

In a well-known article, Carney, Cuddy, and Yap (2010) documented the benefits of “power posing”. In their study, participants (N=42) who were randomly assigned to briefly adopt expansive, powerful postures sought more risk, had higher testosterone levels, and had lower cortisol levels than those assigned to adopt contractive, powerless postures. In their response to a failed replication by Ranehill et al. (2015), Carney, Cuddy, and Yap (2015) reviewed 33 successful studies investigating the effects of expansive vs. contractive posing, focusing on differences between these studies and the failed replication, to identify possible moderators that future studies could explore. But before spending valuable resources on that, it is useful to establish whether the literature that Carney et al. (2015) cited actually suggests that power posing is effective. In this paper we rely on p-curve analysis to answer the following question: Does the literature reviewed by Carney et al. (2015) suggest the existence of an effect once we account for selective reporting? We conclude not. The distribution of p-values from those 33 studies is indistinguishable from what is expected if (1) the average effect size were zero, and (2) selective reporting (of studies and/or analyses) were solely responsible for the significant effects that are published. Although more highly powered future research may find replicable evidence for the purported benefits of power posing (or unexpected detriments), the existing evidence is too weak to justify a search for moderators or to advocate for people to engage in power posing to better their lives.

Eating behaviors

Fattening Fasting: Hungry Grocery Shoppers Buy More Calories, Not More Food

Preordering School Lunch Encourages Better Food Choices by Children

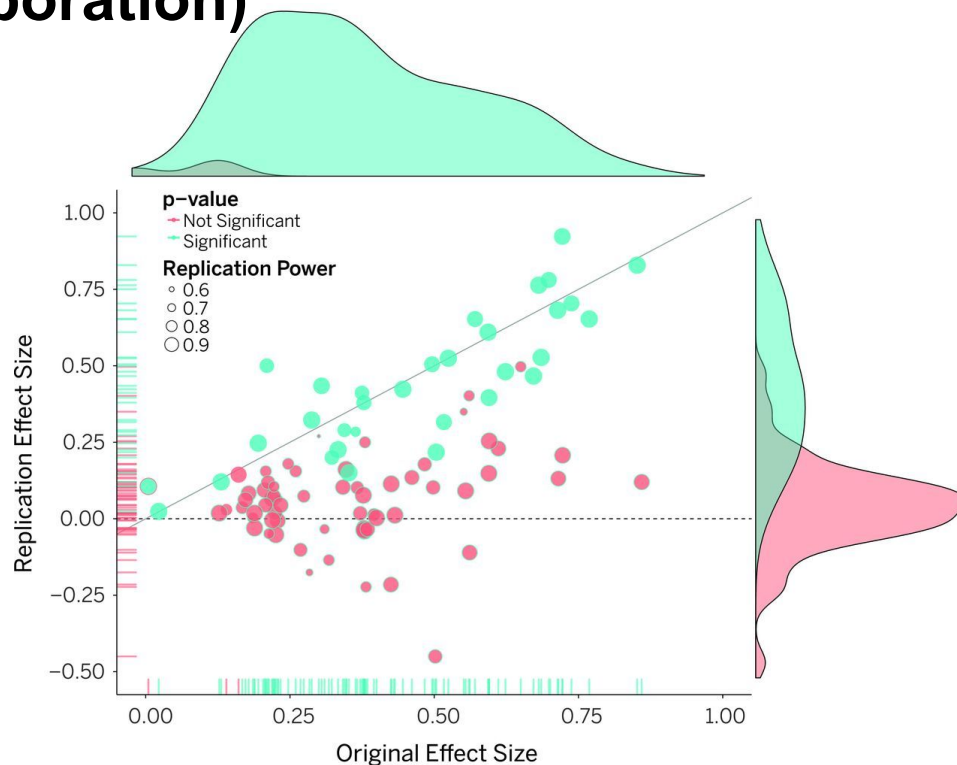
Super Bowls: Serving Bowl Size and Food Consumption

Eating behaviors: P-hacked

Even if you've never heard of Wansink, you're probably familiar with his ideas. His studies, **cited more than 20,000 times**, are about how our environment shapes how we think about food, and what we end up consuming. He's one of the reasons Big Food companies started offering smaller snack packaging, in 100 calorie portions. He once led the USDA committee on dietary guidelines and influenced public policy. He helped Google and the US Army implement programs to encourage healthy eating.

According to **BuzzFeed's Lee**, who obtained Wansink's emails, instead of testing a hypothesis and reporting on whatever findings he came to, Wansink often encouraged his underlings to crunch data in ways that would yield more interesting or desirable results.

Estimating the reproducibility of psychological science (Open Science Collaboration)



P-hacking

"Perhaps the worst fallacy is the kind of self-deception for which psychologist Uri Simonsohn of the University of Pennsylvania and his colleagues have popularized the term *P*-hacking; it is also known as data-dredging, snooping, fishing, significance-chasing and double-dipping. 'P-hacking,' says Simonsohn, 'is trying multiple things until you get the desired result' — even unconsciously. It may be the first statistical term to rate a definition in the online *Urban Dictionary*, where the usage examples are telling: 'That finding seems to have been obtained through *p*-hacking, the authors dropped one of the conditions so that the overall *p*-value would be less than .05,' and 'She is a *p*-hacker, she always monitors data while it is being collected.'"

- *Regina Nuzzo*

P-hacking's family

Preregistering the study and documenting a final analysis plan avoids several pitfalls associated with the recent replication crisis: questionable research practices (John et al., 2012), HARKing -- hypothesizing after results are known (Kerr, 1998), gardens of forking paths (Gelman and Loken, 2014), and p-hacking (Schuemie et al., 2018).

- *Mike Powell, Allison Koenecke, et al.*

Only spooky results get published



“For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results.”

- *Robert Rosenthal*

What happens when publishing $p < 0.05$

One implication is that about 1 in 20 "significant" findings is likely to be a fluke. In practice, the number may be far larger, as scientists often don't publish papers that fail to find a significant result. So, published research is likely to overrepresent the flukey 5 percent. And if the flukey 5 percent are especially interesting, perhaps because of their novel and unexpected findings, then media coverage may exaggerate this overrepresentation even further.

- *Tania Lombrozo*

Preregister hypotheses (Phase 3)

- We can avoid p-hacking by being open about our science!
- Open access to data (if legal and ethical): GitHub, OSF
- Preregistering analyses and hypotheses (e.g. AsPredicted)
 - Pre-establish what analysis you'll run, how many times you'll run it, etc. so we know you aren't cherry picking results
 - Always report the findings from your preregistered analyses, even if they're null results!