# Variation and Randomness

## Understanding Political Numbers

Feb 20, 2019

# Essay 1

Overall great!

Feedback:

- Commit to a measure
- Direction of relationship
- Avoid "proving" and "disproving"
- Contamination of the DV
- Please proofread!

My bad:

- Latent probability vs incidence (e.g. of war)

# Essay 1

Overall great!

Feedback:

- Commit to a measure
- Direction of relationship
- Avoid "proving" and "disproving"
- Contamination of the DV
- Please proofread!

My bad:

- Latent probability vs incidence (e.g. of war)

# R Exercise 1

Installing vs. loading ("librarying") packages

Folder names

Data file name

# Let's talk about papers

Original research question

- Does $X$ affect $Y$?
- What other things might affect $Y$ (at least two other potential explainers)? Use theoretical reasoning

---

Collect data (at least 50 cases)

Write a 12-page paper (not counting graphics)

- What's the question
- Theory and hypotheses
- Explain your data and method

# Let's talk about papers

Original research question

- Does $X$ affect $Y$?
- What other things might affect $Y$ (at least two other potential explainers)? Use theoretical reasoning

Collect data (at least 50 cases)

Write a 12-page paper (not counting graphics)

- What's the question
- Theory and hypotheses
- Explain your data and method

# What to do:

Research question due March 11 (a Monday)

- Variables, hypotheses, and (proposed) data sources
- **Before then:** 15-minute meeting w/ Michael or me
- Extra office hours

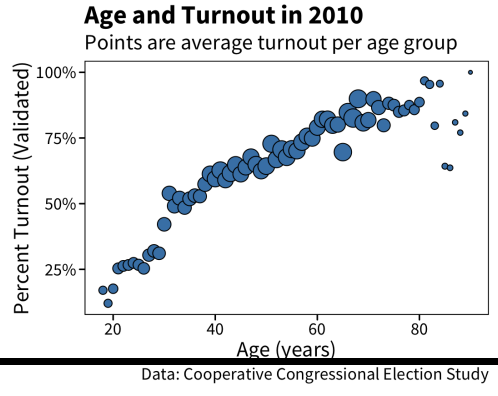
Dataset due after spring break

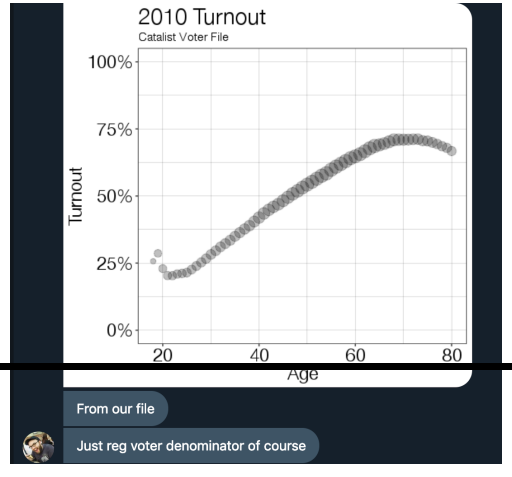
**More info coming soon**

- Watch for email about scheduling meetings! You *must* meet with us
- Formal assignment sheet soon!

https://mikedecr.github.io/p
270/

# Will we ever use this in the real world?



# Where were we?

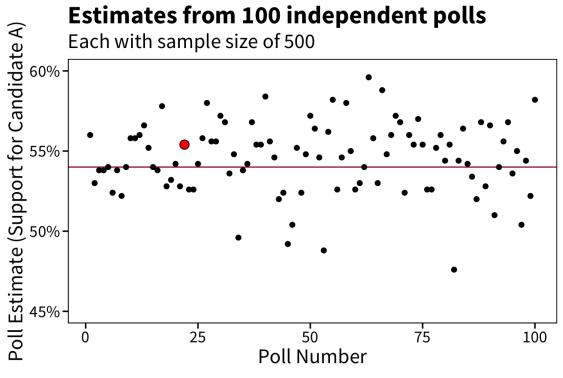
| The "real world" | Your data |
| --- | --- |
| "Population" | Sample |
| Theoretical mean | Sample mean |
| Expectation | Estimate |
| Parameter | Statistic |
| $\mu$ | $x$ |

Remember:

Observed data = Truth + Bias + Noise

# Where were we?

| The "real world" | Your data |
|---|---|
| "Population" | Sample |
| Theoretical mean | Sample mean |
| Expectation | Estimate |
| Parameter | Statistic |



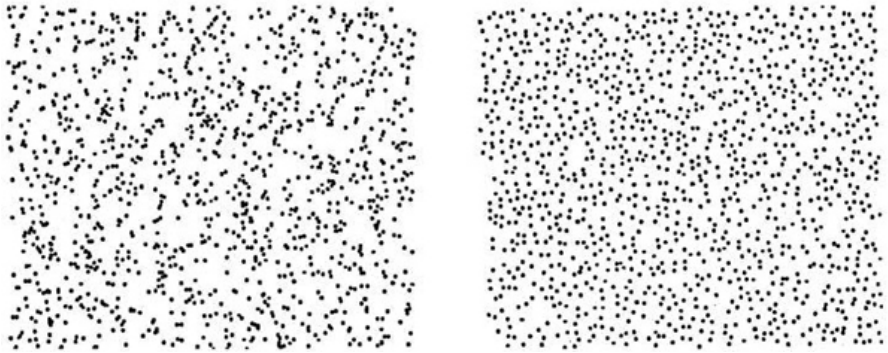

$\mu$ $x$

- In ~~a given sample, I am randomly~~ inaccurate, and that is to be expected.
  - Every time I call someone, there's a 54% chance they support A, so there's randomness
  - When randomness is at play, you get uncertainty
- It looks like it's getting wider, but it isn't.
- That's just random variation

Remember:

Observed data = Truth + Bias + Noise

# Which is random?



# Reality is clumpy

Flip a coin. Probability of heads is $0.5$.

# Reality is clumpy

Flip a coin. Probability of heads is $0.5$.

Flip a coin twice. Probability of two heads in a row is
$0.5 \times 0.5 = 0.25$

# Reality is clumpy

Flip a coin. Probability of heads is $0.5$.

Flip a coin twice. Probability of two heads in a row is
$0.5 \times 0.5 = 0.25$

Flip a coin $n$ times. Probability of $n$ heads in a row is $0.5^n$.

# Reality is clumpy

Flip a coin. Probability of heads is $0.5$.

Flip a coin twice. Probability of two heads in a row is $0.5 \times 0.5 = 0.25$

Flip a coin $n$ times. Probability of $n$ heads in a row is $0.5^n$.

I flip a coin 1000 times. What's the longest sequences of heads I get?

# Reality is clumpy

Flip a coin. Probability of heads is $0.5$.

Flip a coin twice. Probability of two heads in a row is $0.5 \times 0.5 = 0.25$

Flip a coin $n$ times. Probability of $n$ heads in a row is $0.5^n$.

I flip a coin 1000 times. What's the longest sequences of heads I get?

```
# I ran this simulation in R
max(one_thousand_flips$heads_in_a_row)
```

```
## [1] 10
```

Wow! The probability of getting that many heads in a row is $0.5^{10} = 0.0009766$!

# Reality is clumpy

Flip a coin. Probability of heads is $0.5$.

Flip a coin twice. Probability of two heads in a row is $0.5 \times 0.5 = 0.25$

Flip a coin $n$ times. Probability of $n$ heads in a row is $0.5^n$.

I flip a coin 1000 times. What's the longest sequences of heads I get?

```
# I ran this simulation in R
max(one_thousand_flips$heads_in_a_row)
```

```
## [1] 10
```

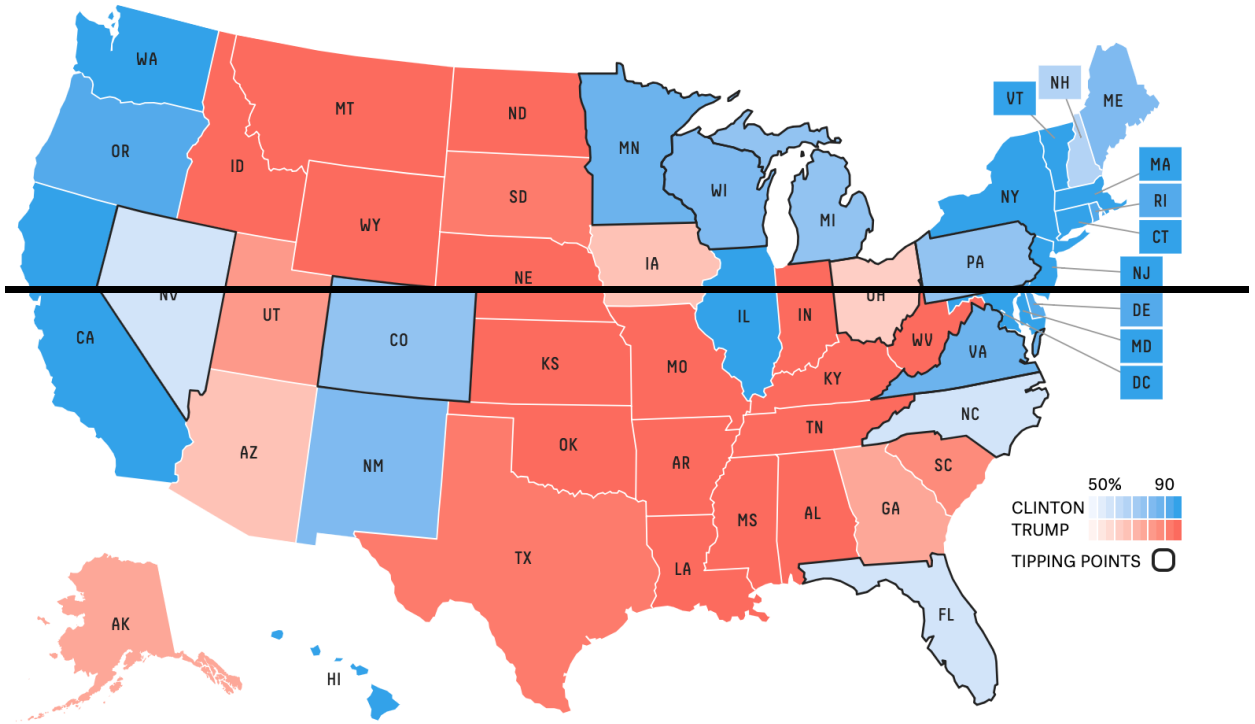
Wow! The probability of getting that many heads in a row is $0.5^{10} = 0.0009766$!

Unlikely things aren't always unusual.

When you have lots of data, unlikely things are common.

# Who will win the presidency?

## Chance of winning

# Time for something amazing

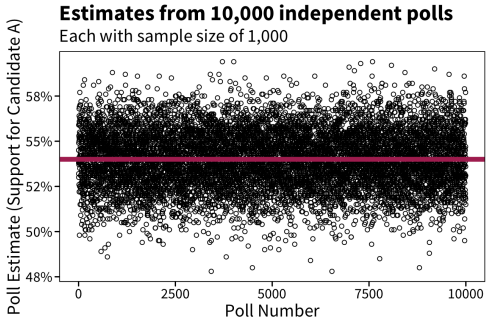
# 10,000 Independent Polls

---

Why 10,000?

- reality is clumpy
- but it's clumpy when you look up close
- when you zoom out really far, clumpiness cancels out (randomly)

# 10,000 Independent Polls




---

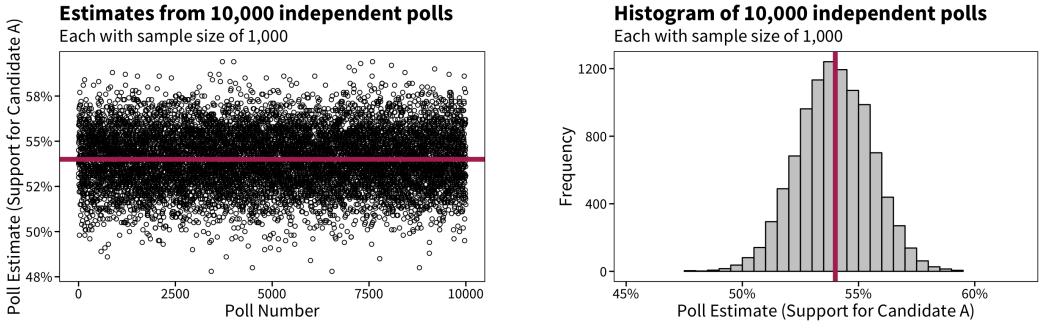
Why 10,000?

- reality is clumpy
- but it's clumpy when you look up close
- when you zoom out really far, clumpiness cancels out (randomly)

bee swarm

- if we were to zoom in, we would see that every poll is different
- Every poll is off a little bit
- but overall, they're close to the truth
- It looks like the farther you get from the truth, the fewer polls are out there

# 10,000 Independent Polls







---

Why 10,000?

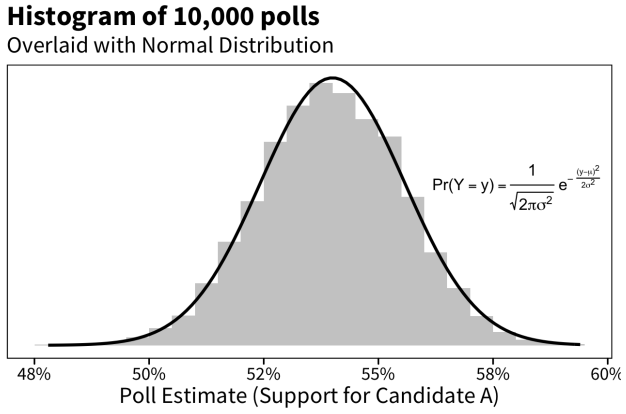
- reality is clumpy
- but it's clumpy when you look up close
- when you zoom out really far, clumpiness cancels out (randomly)

## bee swarm

- if we were to zoom in, we would see that every poll is different
- Every poll is off a little bit
- but overall, they're close to the truth
- It looks like the farther you get from the truth, the fewer polls are out there

# The normal distribution

If some variable $Y$ is affected (at least in part) by an accumulation of random fluctuations, then *the distribution of* $Y$ will be approx. normal




**A "draw" from a normal distribution:**

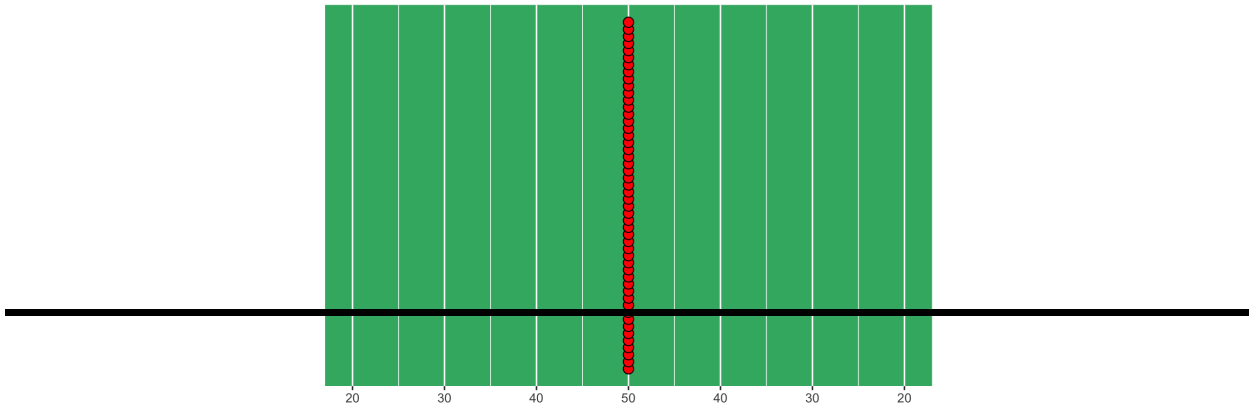
Any *individual* observation of $y$ is just one number, but the probability of observing that value (relative to the mean) is given by the normal distribution

# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations $\rightarrow \infty$)

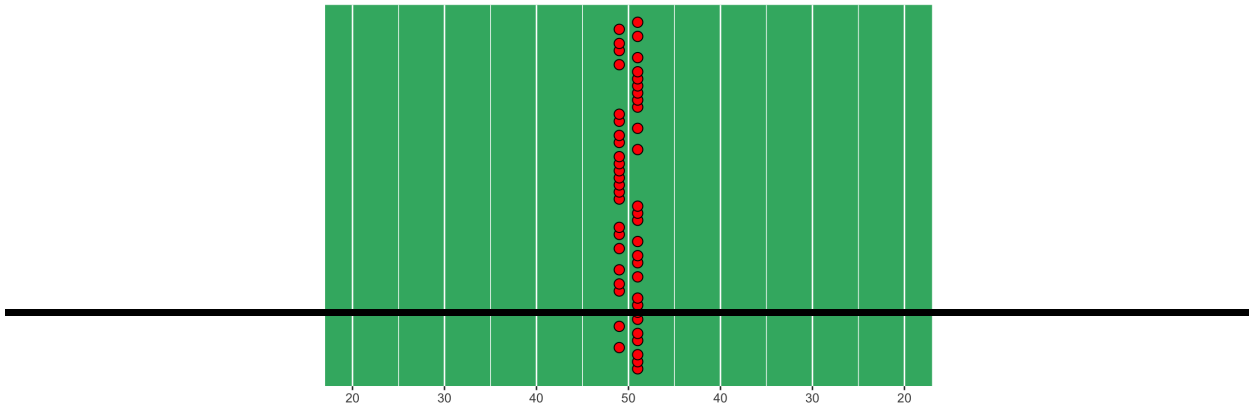
# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations → ∞)



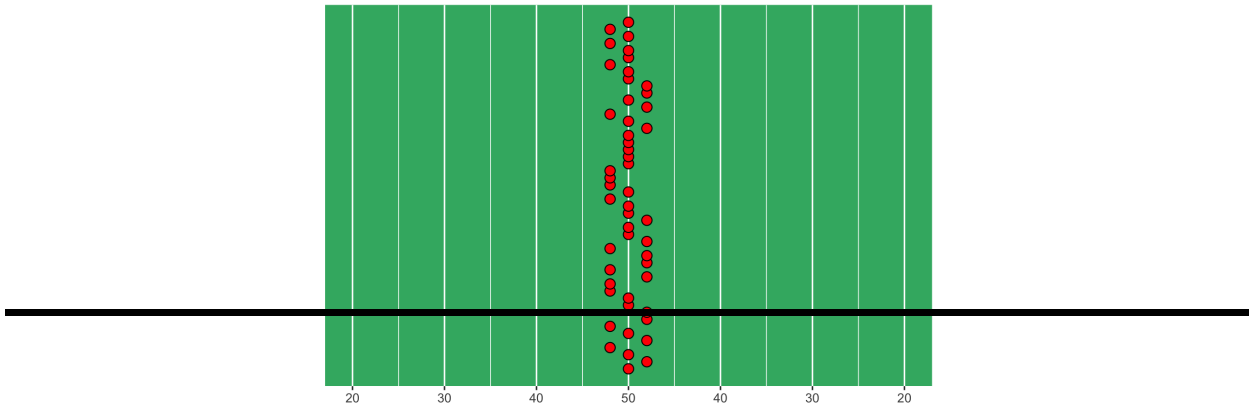
# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations → ∞)



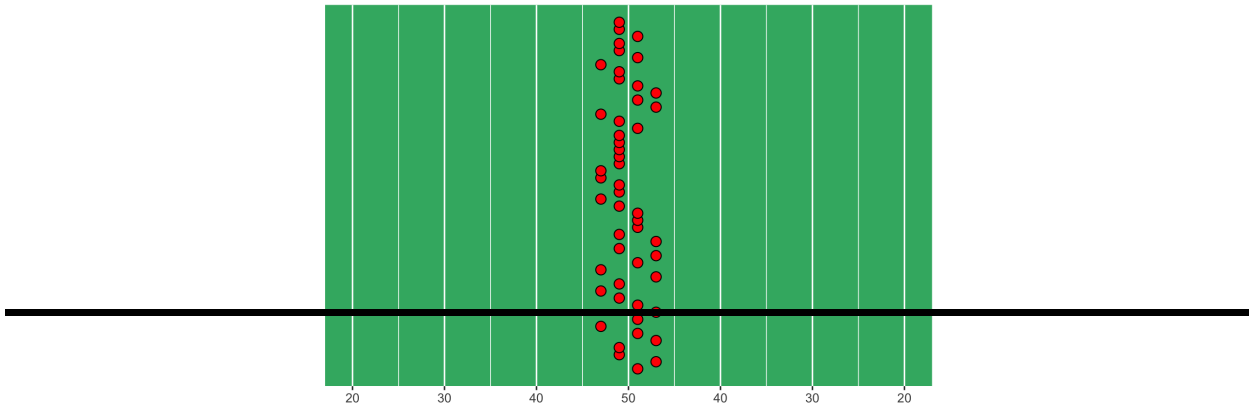
# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations → ∞)



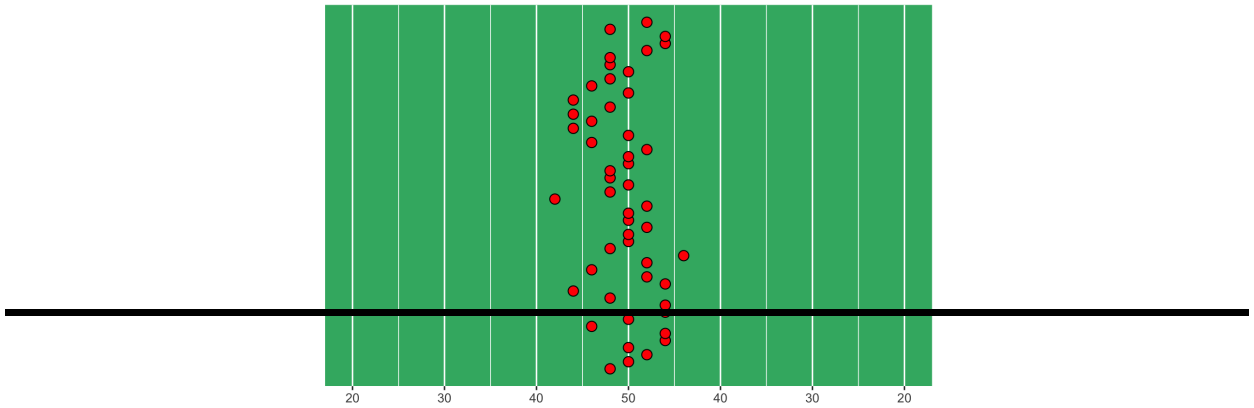
# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations → ∞)



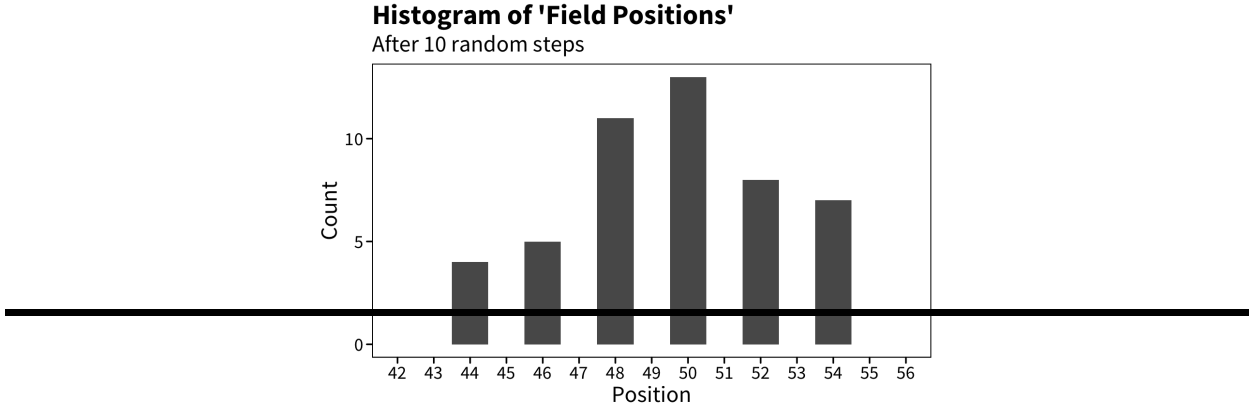
# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations → ∞)



# Accumulating fluctuations

If some outcome is the result of an *accumulation of random fluctuations*, then the outcome is normally distributed (as the number of fluctuations → ∞)



# What do "accumulating fluctuations" have to do with the mean

or, "Why are means normally distributed?"

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

# What do "accumulating fluctuations" have to do with the mean

or, "Why are means normally distributed?"

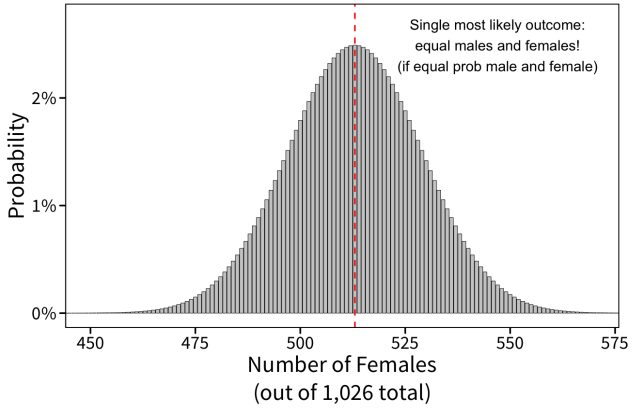
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

~~Summing the data is "adding up fluctuations."~~ **Cool, huh!**

"Order arising from disorder" or just *pure chaos*?

# Most likely babies

# Most likely babies

# How we use the normal distribution in statistics

# "The sampling distribution of $\bar{x}$ is normal"

This means "the sample mean is a *normal draw* from the true mean".

**How accurate is the mean?**

- What's the dispersion in the raw data?
- Calculate the *standard deviation* of data $x$:

$$s(x) = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}}$$

Data $x$ where $\bar{x} = 10$

```
## # A tibble: 100 x 3
##        x `x - xbar` `(x - xbar)^2`
##    <int>      <dbl>          <dbl>
##  1    15          5             25
##  2     1         -9             81
##  3    10          0              0
##  4    14          4             16
##  5    19          9             81
##  6     6         -4             16
##  7    10          0              0
##  8    20         10            100
##  9    16          6             36
## 10    19          9             81
## # … with 90 more rows
```

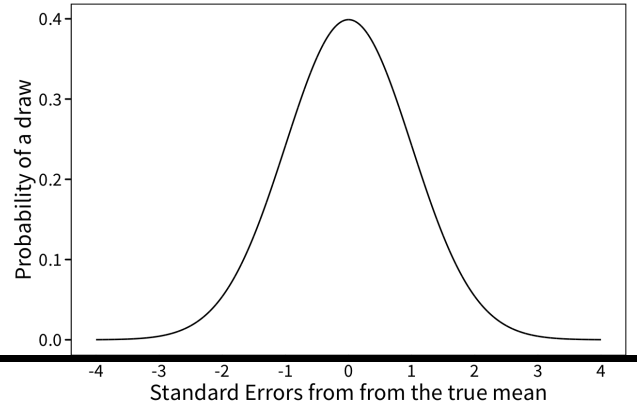
- normal distribution is a single number, but it's more likely to appear in the center of the distribution and less likely to appear in the tail
- We don't know where the true mean is, but we know enough about the properties of the Normal distribution that we can build an informed guess
  - we know the most

# "The sampling distribution of $\bar{x}$ is normal"

**Standard error** is "average error between $\bar{x}$ and the true mean $\mu$"

Standard error of $\bar{x}$:

$$se(\bar{x}) = \frac{s(x)}{\sqrt{n}}$$




Rules for averages:

- ~ 68% of estimates are within $\pm$ 1 std. error from "true" value

# Statistics is about dealing with *uncertainty* in real data

## Here is how most people do statistics

---

Every sample is an *imperfect* representation of the underlying population

Calculate a sample estimate (such as a mean) and its standard error

Most of the time, your estimate is within $\pm$ 2 standard errors from the "true value"

Lots of statistics depend on assumptions about the "distribution" of our data

- a lot of these assumptions are like, assuming that our estimate of the mean is normally distributed
- or assuming that our data are normally distributed

It often feels like this is a tenuous assumption, that could easily be broken

- it is actually hard to break, because it isn't fragile
- The normal distribution is just what happens when you add lots of fluctuations together
- and lots of things are influenced by lots of small fluctuations