

Math 1501 - Elements of Statistics with Applications

1.2 - Data, Sampling, and Variation in Data and Sampling

Two kinds of data

Qualitative data — the result of describing or categorizing. Also called *categorical*.

Two kinds of data

Qualitative data — the result of describing or categorizing. Also called *categorical*.

Quantitative data — always numbers. The result of counting or measuring.

Two kinds of data

Qualitative data — the result of describing or categorizing. Also called *categorical*.

Quantitative data — always numbers. The result of counting or measuring.

Qualitative or quantitative?

Hair color, blood type, the street you live on → ?.

Pulse rate, amount of money, number of students → ?.

Two kinds of data

Qualitative data — the result of describing or categorizing. Also called *categorical*.

Quantitative data — always numbers. The result of counting or measuring.

Qualitative or quantitative?

Hair color, blood type, the street you live on → qualitative.

Pulse rate, amount of money, number of students → quantitative.

Two kinds of data

Qualitative data — the result of describing or categorizing. Also called *categorical*.

Quantitative data — always numbers. The result of counting or measuring.

Qualitative or quantitative?

Hair color, blood type, the street you live on → qualitative.

Pulse rate, amount of money, number of students → quantitative.

There is no *average* hair color. That is not a limitation of our tools — the question itself is meaningless, by definition of 'average'.

Same experiment, two kinds of questions

Class question

A bag holds red, green, and blue balls — far too many to count, but exactly equal numbers of each color. You reach in without looking and draw one ball.

Same experiment, two kinds of questions

Class question

A bag holds red, green, and blue balls — far too many to count, but exactly equal numbers of each color. You reach in without looking and draw one ball.

Question A: Describe the ball by its color — Qualitative or quantitative?

Same experiment, two kinds of questions

Class question

A bag holds red, green, and blue balls — far too many to count, but exactly equal numbers of each color. You reach in without looking and draw one ball.

Question A: Describe the ball by its color — Qualitative or quantitative?

Question B: What is $P(\text{green})$? You can't count every ball in the bag — so what are you actually using to answer this?

Same experiment, two kinds of questions

Class question

A bag holds red, green, and blue balls — far too many to count, but exactly equal numbers of each color. You reach in without looking and draw one ball.

Question A: Describe the ball by its color — Qualitative or quantitative?

Question B: What is $P(\text{green})$? You can't count every ball in the bag — so what are you actually using to answer this?

Same bag, same draw. One question asks you to *describe*; the other asks you to *compute*. The **question** decides what kind of thought process is needed and whether or not a computation makes sense.

The bag has a name

So far: qualitative, quantitative

Population — every individual you could possibly draw from — here, every ball in the bag.

The bag has a name

So far: qualitative, quantitative

Population — every individual you could possibly draw from — here, every ball in the bag.

Parameter — a number that describes the population. Here, $p = \frac{1}{3}$, the *true* proportion of green balls.

The bag has a name

So far: qualitative, quantitative

Population — every individual you could possibly draw from — here, every ball in the bag.

Parameter — a number that describes the population. Here, $p = \frac{1}{3}$, the *true* proportion of green balls.

You will almost never know a parameter exactly. Nobody counts the whole bag.

The bag has a name

So far: qualitative, quantitative

Population — every individual you could possibly draw from — here, every ball in the bag.

Parameter — a number that describes the population. Here, $p = \frac{1}{3}$, the *true* proportion of green balls.

You will almost never know a parameter exactly. Nobody counts the whole bag.

So instead: draw some balls, and describe *those* balls.

Sample — the balls you actually drew.

The bag has a name

So far: qualitative, quantitative

Population — every individual you could possibly draw from — here, every ball in the bag.

Parameter — a number that describes the population. Here, $p = \frac{1}{3}$, the *true* proportion of green balls.

You will almost never know a parameter exactly. Nobody counts the whole bag.

So instead: draw some balls, and describe *those* balls.

Sample — the balls you actually drew.

Statistic — a number that describes the sample — your best guess at the parameter.

One bag, two numbers

So far: qualitative, quantitative, population, parameter, sample, statistic

Worked example

You draw 12 balls: 5 are green.

One bag, two numbers

So far: qualitative, quantitative, population, parameter, sample, statistic

Worked example

You draw 12 balls: 5 are green.

Population proportion (parameter): $p = \frac{1}{3} \approx 0.333$ — true, fixed, usually unknown.

One bag, two numbers

So far: qualitative, quantitative, population, parameter, sample, statistic

Worked example

You draw 12 balls: 5 are green.

Population proportion (parameter): $p = \frac{1}{3} \approx 0.333$ — true, fixed, usually unknown.

Sample proportion (statistic): $\hat{p} = \frac{5}{12} \approx 0.417$ — computed from your draws, changes every time you draw again.

One bag, two numbers

So far: qualitative, quantitative, population, parameter, sample, statistic

Worked example

You draw 12 balls: 5 are green.

Population proportion (parameter): $p = \frac{1}{3} \approx 0.333$ — true, fixed, usually unknown.

Sample proportion (statistic): $\hat{p} = \frac{5}{12} \approx 0.417$ — computed from your draws, changes every time you draw again.

$\hat{p} \neq p$. That gap has a name too — you'll see it again later.

Try it ourselves

In Colab: draw balls from the bag one at a time, and watch the running percentage update live.

Start at 1 ball — 0% or 100% green, nothing in between yet.

Add balls one at a time. Watch the percentage bounce, then settle in toward $p = 1/3$ as the sample grows.

Class question

Sort these into two groups, however feels natural:

- Number of siblings you have
- Your height (inches)
- Number of pets owned
- Time to run a mile (minutes)
- Number of texts sent yesterday
- Weight of a backpack (pounds)

Quantitative data splits again

So far: qualitative, quantitative

These two groups you just built have names:

Discrete — the result of *counting*. Only certain values are possible.

Quantitative data splits again

So far: qualitative, quantitative

These two groups you just built have names:

Discrete — the result of *counting*. Only certain values are possible.

Continuous — the result of *measuring*. Fractions and decimals are possible.

Quantitative data splits again

So far: qualitative, quantitative

These two groups you just built have names:

Discrete — the result of *counting*. Only certain values are possible.

Continuous — the result of *measuring*. Fractions and decimals are possible.

Number of phone calls you got today: 0, 1, 2, 3, ... There is no 2.4.

Length of those calls in minutes: 2.4, 7.5, 11.0 — all fine.

Quantitative data splits again

So far: qualitative, quantitative

These two groups you just built have names:

Discrete — the result of *counting*. Only certain values are possible.

Continuous — the result of *measuring*. Fractions and decimals are possible.

Number of phone calls you got today: 0, 1, 2, 3, ... There is no 2.4.

Length of those calls in minutes: 2.4, 7.5, 11.0 — all fine.

A general shortcut that works most of the time: if the phrase “the number of” fits in front of it, it is discrete.

Showing qualitative data

Pie chart — wedges of a circle, sized by the percent in each category.

Showing qualitative data

Pie chart — wedges of a circle, sized by the percent in each category.

Bar graph — a bar per category, length proportional to its count or percent.

Showing qualitative data

Pie chart — wedges of a circle, sized by the percent in each category.

Bar graph — a bar per category, length proportional to its count or percent.

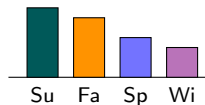
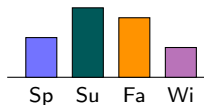
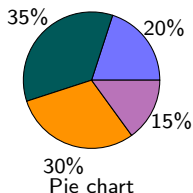
Pareto chart — a bar graph with the bars sorted largest to smallest.

Showing qualitative data

Pie chart — wedges of a circle, sized by the percent in each category.

Bar graph — a bar per category, length proportional to its count or percent.

Pareto chart — a bar graph with the bars sorted largest to smallest.



A pie chart needs a whole to be a pie. If the categories overlap, or if some were left out, the percentages will not total 100% — bar graphs can be more honest... and also not.

Where we left off

We now know how to:

classify a piece of data, and some ways to *visualize* a collection of data.

Where we left off

We now know how to:

classify a piece of data, and some ways to *visualize* a collection of data.

Quick check — qualitative, quantitative, or both split further?

- 1 Favorite pizza topping.
- 2 Number of toppings ordered.
- 3 Time it took the pizza to arrive, in minutes.

Where we left off

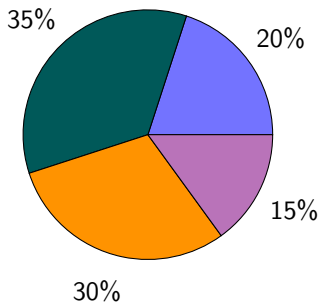
We now know how to:

classify a piece of data, and some ways to *visualize* a collection of data.

Quick check — qualitative, quantitative, or both split further?

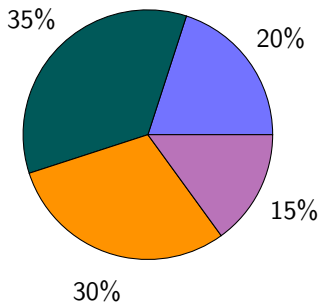
- 1 Favorite pizza topping. Qualitative.
- 2 Number of toppings ordered. Quantitative — discrete.
- 3 Time it took the pizza to arrive, in minutes. Quantitative — continuous.

...and how we picture it — pie chart



Each slice is a fraction of the whole survey — it represents a percentage of the full circle.

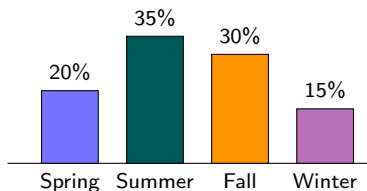
...and how we picture it — pie chart



Each slice is a fraction of the whole survey — it represents a percentage of the full circle.

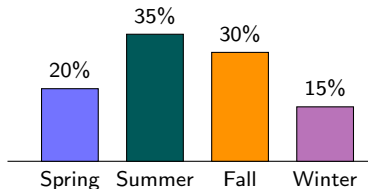
$20\% + 35\% + 30\% + 15\% = 100\%$ — always, because the slices *are* cut from the whole.

...and how we picture it — bar graph



Same numerical data — bar height is just the count or percent for that one category compared to the others.

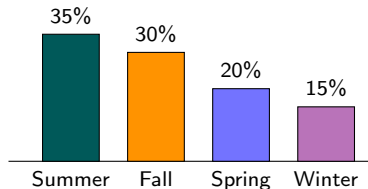
...and how we picture it — bar graph



Same numerical data — bar height is just the count or percent for that one category compared to the others.

Unlike a pie, these bars don't have to add to 100%. Categories could overlap, or some could be left out entirely, and the chart still works.

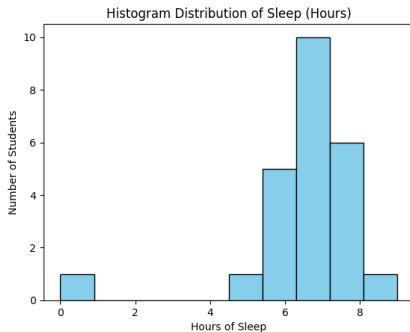
...and how we picture it — Pareto chart



Exact same bars as the bar graph — just sorted largest to smallest, so the biggest category reads first instead of getting lost in alphabetical or arbitrary order.

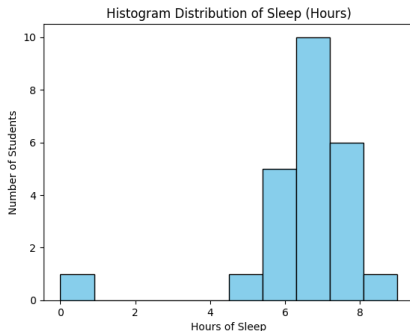
...and how we picture it — Histogram chart

A histogram. Day one, Class survey data plotted, sleep hours from 25 classmates: bars showing *how many students* slept for *how long*.



...and how we picture it — Histogram chart

A histogram. Day one, Class survey data plotted, sleep hours from 25 classmates: bars showing *how many students* slept for *how long*.



That's where most of the class clustered — the most common amount of sleep, read straight off the bar height.

The sample mean, piece by piece

$$x_1$$

each x is one observation — the subscript just says *which one*.

The sample mean, piece by piece

$$x_1 + x_2$$

each x is one observation — the subscript just says *which one*.

The sample mean, piece by piece

$$x_1 + x_2 + x_3$$

each x is one observation — the subscript just says *which one*.

The sample mean, piece by piece

$$x_1 + x_2 + x_3 + \dots$$

each x is one observation — the subscript just says *which one*.

the $\dots$ means “keep doing exactly this,” not something new or mysterious.

The sample mean, piece by piece

$$x_1 + x_2 + x_3 + \cdots + x_n$$

each x is one observation — the subscript just says *which one*.

the $\cdots$ means “keep doing exactly this,” not something new or mysterious.

The sample mean, piece by piece

$$x_1 + x_2 + x_3 + \cdots + x_n$$

each x is one observation — the subscript just says *which one*.

the $\cdots$ means “keep doing exactly this,” not something new or mysterious.

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \bar{x}$$

divide by n , the number of observations you actually have — the result is called $\bar{x}$, the sample mean.

Real numbers, not letters

Day one — your actual sleep data

```
sleep = [7.5, 6, 7.5, 8, 0, 6, 7, 8, 7, 6, 6.5, 6, 7,  
         9, 7, 5, 6.5, 8, 7, 7.5, 7, 6, 7, 7]
```

Real numbers, not letters

Day one — your actual sleep data

```
sleep = [7.5, 6, 7.5, 8, 0, 6, 7, 8, 7, 6, 6.5, 6, 7,  
         9, 7, 5, 6.5, 8, 7, 7.5, 7, 6, 7, 7]
```

$n = 24$ — count the entries in the list. That's how many students were in the room that day.

Real numbers, not letters

Day one — your actual sleep data

```
sleep = [7.5, 6, 7.5, 8, 0, 6, 7, 8, 7, 6, 6.5, 6, 7,  
         9, 7, 5, 6.5, 8, 7, 7.5, 7, 6, 7, 7]
```

$n = 24$ — count the entries in the list. That's how many students were in the room that day.

$$x_1 + x_2 + \cdots + x_{24} = 159.5$$

Real numbers, not letters

Day one — your actual sleep data

```
sleep = [7.5, 6, 7.5, 8, 0, 6, 7, 8, 7, 6, 6.5, 6, 7,  
         9, 7, 5, 6.5, 8, 7, 7.5, 7, 6, 7, 7]
```

$n = 24$ — count the entries in the list. That's how many students were in the room that day.

$$x_1 + x_2 + \cdots + x_{24} = 159.5$$

$\frac{159.5}{24} = \bar{x} \approx 6.65$ hours — one more decimal place than the data had, same rounding rule as always.

Same class, a different day

Day one: $n = 24$, $\bar{x} \approx 6.65$ hours.

Same class, a different day

Day one: $n = 24$, $\bar{x} \approx 6.65$ hours.

Class three had more students in the room — $n = 35$ instead of 24.

Same class, a different day

Day one: $n = 24$, $\bar{x} \approx 6.65$ hours.

Class three had more students in the room — $n = 35$ instead of 24.

Recompute with 35 responses and we will almost certainly get a *different* $\bar{x}$ — not because anyone's sleep habits changed drastically, but because a different sample count gives a different estimate.

Same class, a different day

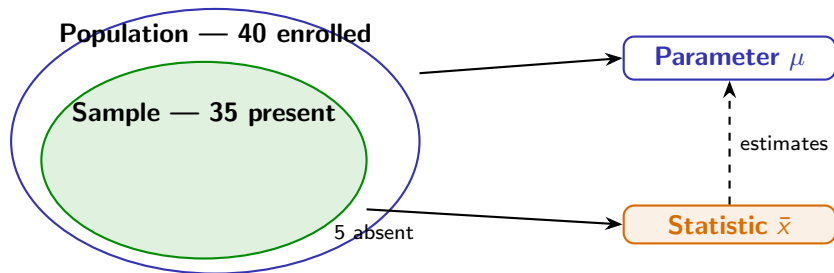
Day one: $n = 24$, $\bar{x} \approx 6.65$ hours.

Class three had more students in the room — $n = 35$ instead of 24.

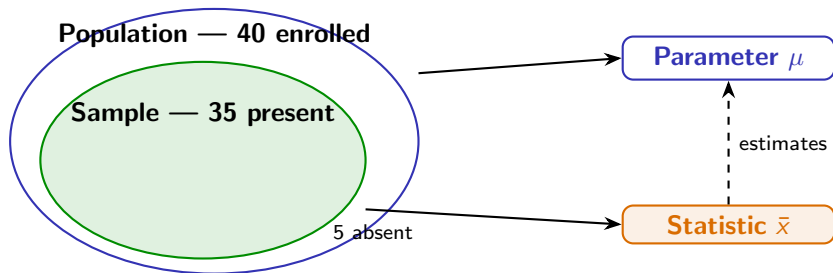
Recompute with 35 responses and we will almost certainly get a *different* $\bar{x}$ — not because anyone's sleep habits changed drastically, but because a different sample count gives a different estimate.

$\bar{x}$ is never final. It is always a statistic — your best estimate from whoever happened to be in the room and chose to respond. Only the true average across every student in the class, μ , would be fixed — and we will never know it exactly unless the entire class is present and responds again.

The picture behind the numbers



The picture behind the numbers



Even 35 out of 40 is still a *sample*. $\bar{x}$ is still a *statistic* — an estimate of μ , not μ itself. Only all 40, answering, would give you the parameter.

Why sample at all?

Measuring an entire population — every Brooklyn College student, every registered voter — is almost always too expensive or simply impossible.

Why sample at all?

Measuring an entire population — every Brooklyn College student, every registered voter — is almost always too expensive or simply impossible.

So we measure a smaller group instead, and use it to stand in for the whole.

Why sample at all?

Measuring an entire population — every Brooklyn College student, every registered voter — is almost always too expensive or simply impossible.

So we measure a smaller group instead, and use it to stand in for the whole.

But *how* that smaller group gets chosen decides whether it actually can stand in for the population — or quietly misrepresents it. That is the whole reason the sampling methods have names.

Four ways to sample at random, and one that is not

So far: qualitative, quantitative, discrete, continuous

Simple random — every group of n individuals is equally likely to be chosen.

Four ways to sample at random, and one that is not

So far: qualitative, quantitative, discrete, continuous

Simple random — every group of n individuals is equally likely to be chosen.

Stratified — split into groups, take a proportionate number from each group.

Four ways to sample at random, and one that is not

So far: qualitative, quantitative, discrete, continuous

Simple random — every group of n individuals is equally likely to be chosen.

Stratified — split into groups, take a proportionate number from each group.

Cluster — split into groups, pick whole groups at random, take everyone in them.

Four ways to sample at random, and one that is not

So far: qualitative, quantitative, discrete, continuous

Simple random — every group of n individuals is equally likely to be chosen.

Stratified — split into groups, take a proportionate number from each group.

Cluster — split into groups, pick whole groups at random, take everyone in them.

Systematic — pick a random starting point, then take every k th one.

Four ways to sample at random, and one that is not

So far: qualitative, quantitative, discrete, continuous

Simple random — every group of n individuals is equally likely to be chosen.

Stratified — split into groups, take a proportionate number from each group.

Cluster — split into groups, pick whole groups at random, take everyone in them.

Systematic — pick a random starting point, then take every k th one.

Convenience — use whoever is easy to reach. *Not random.*

Same question, five different samples

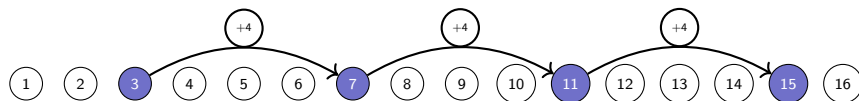
Simple random	every group of n equally likely to be chosen
Stratified	proportionate number taken from <i>every</i> group
Cluster	everyone taken from a few randomly chosen groups
Systematic	random start, then every k th one (explained next slide)
Convenience	whoever is easy to reach — <i>not</i> random

Worked example

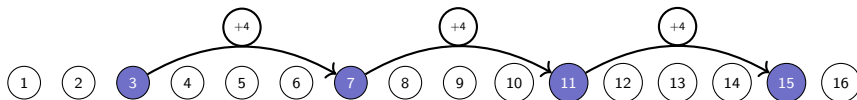
A study wants the average tuition San Jose State students pay per semester.

Same study, five different ways someone could have collected the data.
Which method is which?

Every k th one

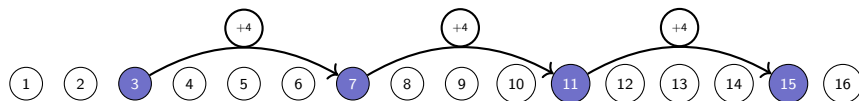


Every k th one



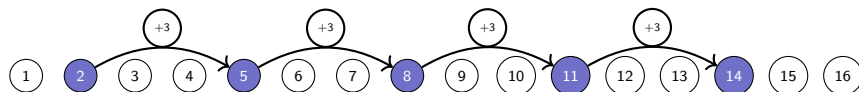
The starting point — position 3 — was picked at random.

Every k th one



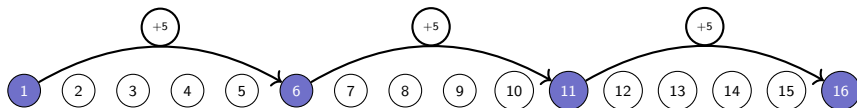
The starting point — position 3 — was picked at random. After that, every step is fixed: +4, every time. No more randomness once k is set.

Every k th one



A different random start, a different step size — same rule: pick k , pick a start, then count forward.

Every k th one



And again, with $k = 5$. The pattern never depends on the numbers chosen — only on picking the random start once, before anything else happens.

Same question, five different samples

Simple random	every group of n equally likely to be chosen
Stratified	proportionate number taken from every group
Cluster	everyone taken from a few randomly chosen groups
Systematic	random start, then every k th one
Convenience	whoever is easy to reach — <i>not</i> random

A study wants the average tuition San Jose State students pay per semester.

- 1 100 students sorted by year; 25 taken from each.
- 2 Random start, then every 50th student, alphabetically.
- 3 Every student equally likely, at every stage.
- 4 Two of the four years picked at random; everyone in them surveyed.
- 5 The first 100 students an assistant happens to meet.

Same question, five different samples

Simple random	every group of n equally likely to be chosen
Stratified	proportionate number taken from every group
Cluster	everyone taken from a few randomly chosen groups
Systematic	random start, then every k th one
Convenience	whoever is easy to reach — <i>not</i> random

A study wants the average tuition San Jose State students pay per semester.

- 1 100 students sorted by year; 25 taken from each. **Stratified.**
- 2 Random start, then every 50th student, alphabetically. **Systematic.**
- 3 Every student equally likely, at every stage. **Simple random.**
- 4 Two of the four years picked at random; everyone in them surveyed. **Cluster.**
- 5 The first 100 students an assistant happens to meet. **Convenience.**

Three things that can introduce error

So far: qualitative, quantitative, discrete, continuous, simple random, stratified, cluster, systematic, convenience

Sampling error — the natural variation from measuring a finite sample instead of everyone or every possible outcome.

Three things that can introduce error

So far: qualitative, quantitative, discrete, continuous, simple random, stratified, cluster, systematic, convenience

Sampling error — the natural variation from measuring a finite sample instead of everyone or every possible outcome.

Nonsampling error — everything else: broken instruments, typos, bad study design.

Three things that can introduce error

So far: qualitative, quantitative, discrete, continuous, simple random, stratified, cluster, systematic, convenience

Sampling error — the natural variation from measuring a finite sample instead of everyone or every possible outcome.

Nonsampling error — everything else: broken instruments, typos, bad study design.

Sampling bias — some members of the population were less likely to be chosen than others.

Variation will naturally exist in populations

Eight 16-ounce cans, measured:

15.8, 16.1, 15.2, 14.8, 15.8, 15.9, 16.0, 15.5. Nothing is broken — measurement varies, and filling machines vary.

Variation will naturally exist in populations

Eight 16-ounce cans, measured:

15.8, 16.1, 15.2, 14.8, 15.8, 15.9, 16.0, 15.5. Nothing is broken — measurement varies, and filling machines vary.

Sampling error can be made smaller with a larger sample. It never reaches zero.

Sampling bias cannot be fixed by collecting more data — a bigger biased sample gives a more confident biased answer.

See video.

Trustworthy data collection?

So far: sampling error, nonsampling error, sampling bias

Three real reasons a study's data collection might not be trustworthy:

Self-selected — online and call-in polls — only people who cared enough answered.

Trustworthy data collection?

So far: sampling error, nonsampling error, sampling bias

Three real reasons a study's data collection might not be trustworthy:

Self-selected — online and call-in polls — only people who cared enough answered.

Undue influence — the question was worded to get a particular answer.

Trustworthy data collection?

So far: sampling error, nonsampling error, sampling bias

Three real reasons a study's data collection might not be trustworthy:

Self-selected — online and call-in polls — only people who cared enough answered.

Undue influence — the question was worded to get a particular answer.

Non-response — the people who refused to answer may be exactly the interesting ones.

Trustworthy data collection? cont.

So far: self-selected, undue influence, non-response

Two more — these are about what the data *means*, not how it was collected:

Causality — two things moving together does not mean one caused the other.

Trustworthy data collection? cont.

So far: self-selected, undue influence, non-response

Two more — these are about what the data *means*, not how it was collected:

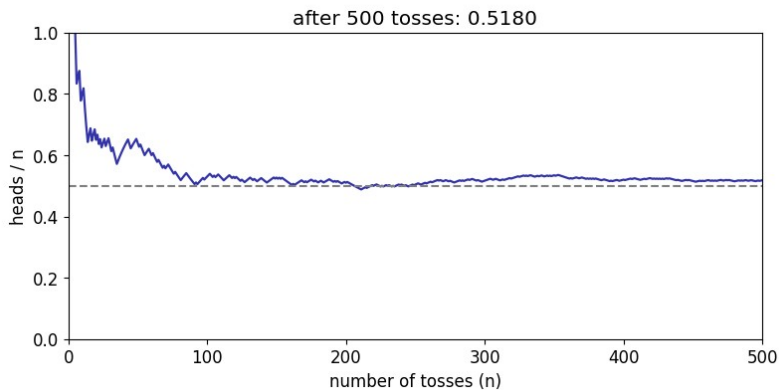
Causality — two things moving together does not mean one caused the other.

Who paid — and what did they need the study to say?

Do not assume a study is bad just because one of these applies. Evaluate it on its merits.

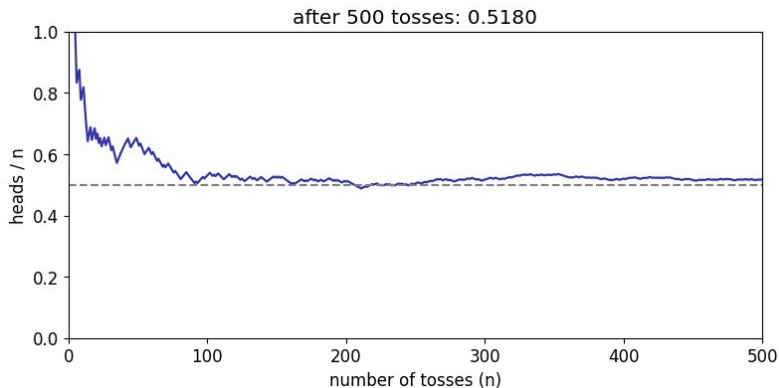
Sampling error now named

Callback to 1.1 — “What we just watched”



Sampling error now named

Callback to 1.1 — “What we just watched”



That wobble on the left, settling down on the right — that *was* sampling error. Now it has a name.