

Math 1501 - Elements of Statistics with Applications

1.1 - Definitions of Statistics, Probability, and Key Terms

Chapter 1: Sampling and Data

1.1 Definitions of Statistics, Probability, and Key Terms

1.2 Data, Sampling, and Variation

1.3 Frequency Tables and Levels of Measurement

1.4 Experimental Design and Ethics

By the end of 1.1 you should be able to answer:

- Who are we studying, and who did we actually measure?
- What is the difference between describing data and concluding from it?
- Which numbers can we compute, and which would be nonsense?

Let's collect some data — and then let Python draw it

On average, to the nearest half hour, how many hours do you sleep per night?

We will type every answer in the room into a list and get a picture back.

```
sleep = [5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9]
```

[Google Colab Link](#)

Let's collect some data — and then let Python draw it

On average, to the nearest half hour, how many hours do you sleep per night?

We will type every answer in the room into a list and get a picture back.

```
sleep = [5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9]
```

That is a *list* in Python. It is how data gets typed in and stored.

This is not a coding course. Python is a tool we point at data, the same way we point a calculator at arithmetic. [Google Colab Link](#)

The two halves of the subject

Statistics — the science of collecting, analyzing, interpreting, and presenting data.

The two halves of the subject

Statistics — the science of collecting, analyzing, interpreting, and presenting data.

Descriptive statistics — organizing and summarizing data – with a graph, or with a number such as an average.

The two halves of the subject

Statistics — the science of collecting, analyzing, interpreting, and presenting data.

Descriptive statistics — organizing and summarizing data – with a graph, or with a number such as an average.

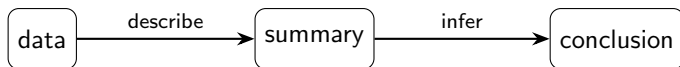
Inferential statistics — using data from a small group to draw a conclusion about a much larger group.

The two halves of the subject

Statistics — the science of collecting, analyzing, interpreting, and presenting data.

Descriptive statistics — organizing and summarizing data – with a graph, or with a number such as an average.

Inferential statistics — using data from a small group to draw a conclusion about a much larger group.



Why probability shows up in a statistics course

Toss a fair coin 4 times. You may well get 3 heads. Nothing is wrong.

Toss it 4,000 times and the fraction of heads sits very close to one half.

Karl Pearson (Mathematician) actually did this. He tossed a coin 24,000 times and got 12,012 heads.

What that fraction is actually saying

$$\frac{12,012}{24,000} = 0.5005$$

What that fraction is actually saying

$$\frac{12,012}{24,000} = 0.5005$$

$$\frac{\text{the part we counted}}{\text{the whole group}}$$

What that fraction is actually saying

$$\frac{12,012}{24,000} = 0.5005$$

$$\frac{\text{the part we counted}}{\text{the whole group}}$$

The two extremes tell you everything:

$$\frac{0}{24,000} = 0 \quad \text{never happened}$$

$$\frac{24,000}{24,000} = 1 \quad \text{happened every time}$$

What that fraction is actually saying

$$\frac{12,012}{24,000} = 0.5005$$

$$\frac{\text{the part we counted}}{\text{the whole group}}$$

The two extremes tell you everything:

$$\frac{0}{24,000} = 0 \quad \text{never happened}$$

$$\frac{24,000}{24,000} = 1 \quad \text{happened every time}$$

The part can never exceed the whole. So this number always lands between 0 and 1.

Naming what just happened: $P(A)$

So far: statistics, descriptive, inferential

Event — one outcome, or a group of outcomes, from a random experiment. Written with a capital letter: A , B , $\dots$

Naming what just happened: $P(A)$

So far: statistics, descriptive, inferential

Event — one outcome, or a group of outcomes, from a random experiment. Written with a capital letter: A , B , ...

Sample space — the set of every possible outcome. Written S .

Naming what just happened: $P(A)$

So far: statistics, descriptive, inferential

Event — one outcome, or a group of outcomes, from a random experiment. Written with a capital letter: A , B , ...

Sample space — the set of every possible outcome. Written S .

$P(A)$ — read “the probability of A .” The same part-over-whole fraction from a moment ago, now with a name for each piece.

Naming what just happened: $P(A)$

So far: statistics, descriptive, inferential

Event — one outcome, or a group of outcomes, from a random experiment. Written with a capital letter: A , B , ...

Sample space — the set of every possible outcome. Written S .

$P(A)$ — read “the probability of A .” The same part-over-whole fraction from a moment ago, now with a name for each piece.

$$P(A) = \frac{\text{outcomes that count as } A}{\text{outcomes in the sample space } S}$$

Naming what just happened: $P(A)$

So far: statistics, descriptive, inferential

Event — one outcome, or a group of outcomes, from a random experiment. Written with a capital letter: A , B , $\dots$

Sample space — the set of every possible outcome. Written S .

$P(A)$ — read “the probability of A .” The same part-over-whole fraction from a moment ago, now with a name for each piece.

$$P(A) = \frac{\text{outcomes that count as } A}{\text{outcomes in the sample space } S}$$

Roll a die. Let $A =$ “you roll a 2.” $S = \{1, 2, 3, 4, 5, 6\}$. $P(A) = \frac{1}{6}$.

Three rules, and that is all

Definition

Probability is a number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur.

Three rules, and that is all

Definition

Probability is a number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur.

1. never negative, never more than certain

$$0 \leq P(A) \leq 1$$

Three rules, and that is all

Definition

Probability is a number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur.

1. never negative, never more than certain
2. something in the sample space S has to happen

$$0 \leq P(A) \leq 1$$

$$P(S) = 1$$

Three rules, and that is all

Definition

Probability is a number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur.

1. never negative, never more than certain
2. something in the sample space S has to happen
3. both events can't happen simultaneously?
sum

$$0 \leq P(A) \leq 1$$

$$P(S) = 1$$

$$P(A \text{ or } B) = P(A) + P(B)$$

Three rules, and that is all

Definition

Probability is a number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur.

1. never negative, never more than certain $0 \leq P(A) \leq 1$
2. something in the sample space S has to happen $P(S) = 1$
3. both events can't happen simultaneously?
sum $P(A \text{ or } B) = P(A) + P(B)$

3. Explained: Roll a die. A 2 and a 5 cannot both happen, so $P(2 \text{ or } 5) = \frac{1}{6} + \frac{1}{6}$.

Note: an event is a collection of outcomes — that is why we can add them.

Let's watch it happen

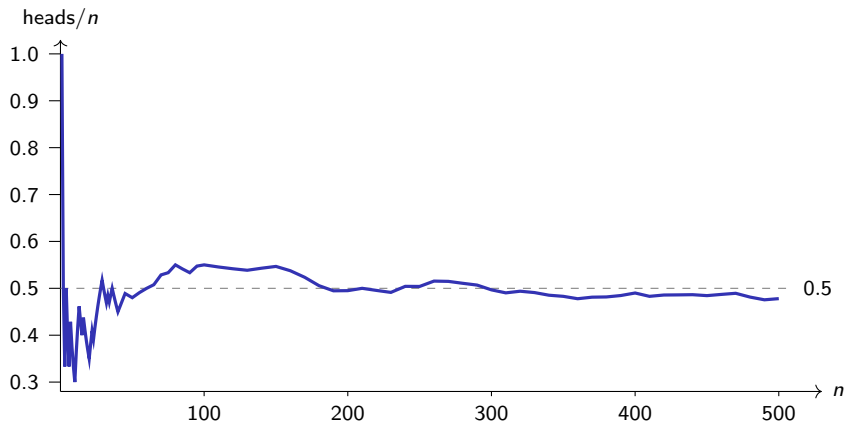
Over to Colab. We toss a coin over and over, and after every single toss we recompute

$$\frac{\text{heads so far}}{\text{tosses so far}}$$

and plot it. The picture builds itself in front of you as n grows.

Then a slider, so you can choose n yourself and run it again.

What we just watched



Who we study, and who we measure

So far: statistics, descriptive, inferential, event, sample space, probability

Population — every person, thing, or object under study.

Who we study, and who we measure

So far: statistics, descriptive, inferential, event, sample space, probability

Population — every person, thing, or object under study.

Sample — the portion of the population we actually measure.

Who we study, and who we measure

So far: statistics, descriptive, inferential, event, sample space, probability

Population — every person, thing, or object under study.

Sample — the portion of the population we actually measure.

Data — the values we obtain from the sample. A single value is called a *datum*.

Who we study, and who we measure

So far: statistics, descriptive, inferential, event, sample space, probability

Population — every person, thing, or object under study.

Sample — the portion of the population we actually measure.

Data — the values we obtain from the sample. A single value is called a *datum*.

Representative sample — a sample that carries the characteristics of the population.

Who we study, and who we measure

So far: statistics, descriptive, inferential, event, sample space, probability

Population — every person, thing, or object under study.

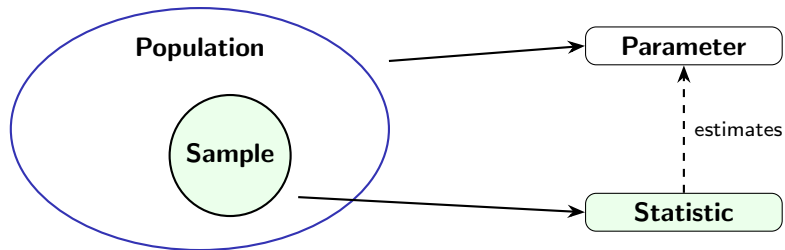
Sample — the portion of the population we actually measure.

Data — the values we obtain from the sample. A single value is called a *datum*.

Representative sample — a sample that carries the characteristics of the population.

Measuring an entire population usually costs too much or is impossible. So we sample.

The whole course in one picture



Two numbers that look alike and are not

So far: population, sample, data, representative sample

Parameter — a number describing the whole *population*. Usually unknown.

Two numbers that look alike and are not

So far: population, sample, data, representative sample

Parameter — a number describing the whole *population*. Usually unknown.

Statistic — a number describing the *sample*. Always computable.

Two numbers that look alike and are not

So far: population, sample, data, representative sample

Parameter — a number describing the whole *population*. Usually unknown.

Statistic — a number describing the *sample*. Always computable.

Population described by a **parameter**

Sample described by a **statistic**

Two numbers that look alike and are not

So far: population, sample, data, representative sample

Parameter — a number describing the whole *population*. Usually unknown.

Statistic — a number describing the *sample*. Always computable.

Population described by a **parameter**

Sample described by a **statistic**

A statistic is our estimate of a parameter. How good that estimate is depends entirely on how well the sample represents the population.

The symbols, so they are never a surprise

	Population (parameter)	Sample (statistic)
size	N	n
mean	μ	$\bar{x}$
standard deviation	σ	s
variance	σ^2	s^2
proportion	p	$\hat{p}$

The symbols, so they are never a surprise

	Population (parameter)	Sample (statistic)
size	N	n
mean	μ	$\bar{x}$
standard deviation	σ	s
variance	σ^2	s^2
proportion	p	$\hat{p}$

You are not computing any of these today.

Greek letters live in the population. Latin letters live in the sample. That single habit will keep you oriented all semester.

Variables: what we record about each member

So far: parameter, statistic

Variable — a characteristic measured for each member of a population.

Written X , Y , ...

Variables: what we record about each member

So far: parameter, statistic

Variable — a characteristic measured for each member of a population.

Written X , Y , ...

Numerical variable — takes values that are numbers with units, such as weight or time.

Variables: what we record about each member

So far: parameter, statistic

Variable — a characteristic measured for each member of a population.
Written X , Y , ...

Numerical variable — takes values that are numbers with units, such as weight or time.

Categorical variable — places each member into a category, such as party affiliation.

Variables: what we record about each member

So far: parameter, statistic

Variable — a characteristic measured for each member of a population.
Written X , Y , ...

Numerical variable — takes values that are numbers with units, such as weight or time.

Categorical variable — places each member into a category, such as party affiliation.

Let X = points earned by a student. Let Y = a person's party affiliation.

We can average the values of X . An “average party affiliation” means nothing.

Two statistics we will use all semester

So far: variable, numerical, categorical

Mean — add the values, divide by how many there are.

Two statistics we will use all semester

So far: variable, numerical, categorical

Mean — add the values, divide by how many there are.

Proportion — the number of members in a category, divided by the total.

Two statistics we will use all semester

So far: variable, numerical, categorical

Mean — add the values, divide by how many there are.

Proportion — the number of members in a category, divided by the total.

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

$$\hat{p} = \frac{\text{number in the category}}{\text{total number}}$$

Words we now own

Population	everyone or everything under study
Sample	the portion actually measured
Parameter	a number describing the population (μ , σ , p)
Statistic	a number describing the sample ($\bar{x}$, s , $\hat{p}$)
Variable	a characteristic measured for each member
Probability	a number in $[0, 1]$ measuring likelihood

Six words. Everything else in 1.1 was built out of these.

Question 1

We want to know the mean amount of money a first-year Brooklyn College student spends on transportation each month.

We randomly survey 120 first-year students. Three of them spent \$88, \$132, and \$145.

Identify: population, sample, parameter, statistic, variable, data.

Question 2

Parameter or statistic? Say which, and say why.

- 1 The mean age of all 15,000 undergraduates at this college.
- 2 The mean age of the 30 students in this class.
- 3 The proportion of all New York City residents who own a car.
- 4 The proportion of 500 surveyed New Yorkers who own a car.