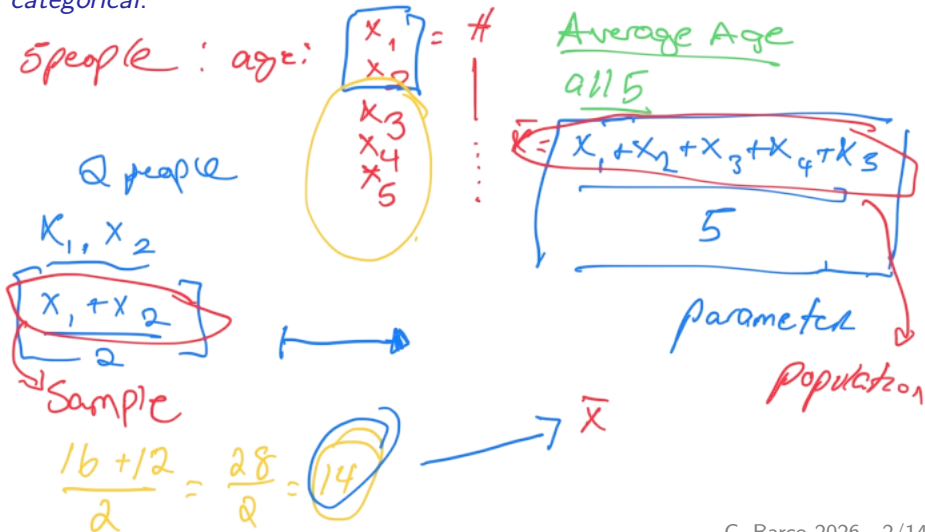


Math 1501 - Elements of Statistics with Applications

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)

Siblings whole
pet
of texts

Fractions
height in inches
Time
weight

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.

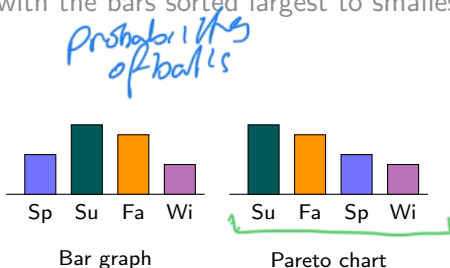
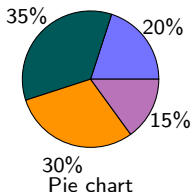
Pareto chart — a bar graph with the bars sorted largest to smallest.

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.

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.

$$n=2$$

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.

k = Choice of the sample