Bootstrapping & confidence intervals

*the **infer** workflow, one verb at a time*

specify()
generate()
calculate()
get_confidence_interval()

resampling with replacement

CI

The mean weight of a bag of chocolate-covered almonds

The idea

You only ever get one sample. Bootstrapping asks what other samples like it would have looked like.

Resample Your Sample

- Treat your sample as a stand-in for the population.
- Draw from it **with replacement**, same size, over and over.
- Each resample gives another value of the statistic.
- Their spread estimates the **standard error** you could not otherwise see.

specify()

Name the variable you are studying. Nothing is computed yet: this only records **meta-data** about the question.

Name Your Variable

Give it a response variable, or a formula y ~ x for two variables.

specify()

Example

the data frame

```
almonds_sample_100 |>
  specify(response = weight)
```

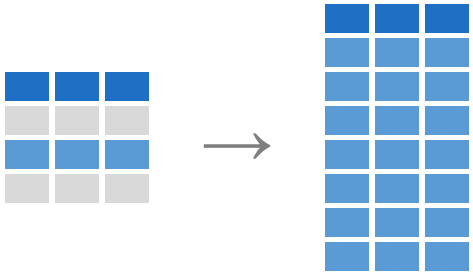
the verb (function)

which variable you are studying

generate()

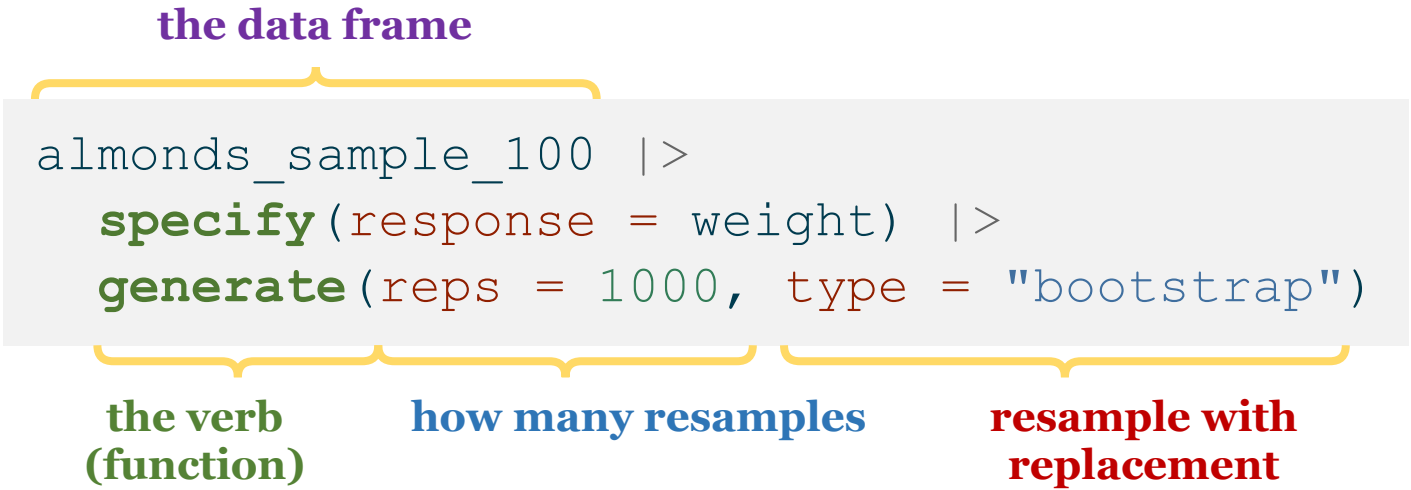
Build the resamples. **type** = **“bootstrap”** draws from your sample with replacement, as many times as **reps** says.

Resample A Thousand Times



generate()

Example



calculate()

Reduce every resample to one number. **stat** names which statistic: a mean, a proportion, a difference, a slope.

One Number Per Resample

The result has one row per replicate, and that is the bootstrap distribution.

calculate()

the bootstrap distribution, saved for later

```
bootstrap_means <- almonds_sample_100 |>
  specify(response = weight) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean")
```

the verb (function)

which statistic, for each resample

get_confidence_interval()

Take the middle 95% of the bootstrap distribution and report its two endpoints.

Example

the bootstrap distribution

```
percentile_ci <- bootstrap_means |>
  get_confidence_interval(level = 0.95, type = "percentile")
```

the verb (function)

how confident

cut at the 2.5th and 97.5th percentiles

visualize()

Plot the bootstrap distribution and shade the interval on top of it.

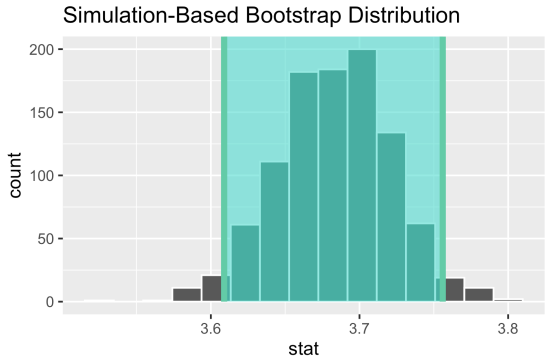
the bootstrap distribution

```
visualize(bootstrap_means) +
  shade_confidence_interval(endpoints = percentile_ci)
```

shade the interval

the two endpoints you just computed

Example



What it means

The sentence to write, and the one to avoid.

Interpreting A 95% CI

- Say: we are 95% **confident** the population mean lies in this interval.
- The confidence is in the **procedure**, not in this one interval.
- Repeat the study many times and about 95% of the intervals would cover it.
- Do **not** say there is a 95% chance the parameter is in this interval.