Sampling

the sampling distribution

*simulating repeated samples with **moderndive***

`p-hat`

rep_slice_sample()
summarize()
geom_histogram()

A bowl of red and white balls, sampled over and over

Four words

Keep these straight and the rest of the chapter follows.

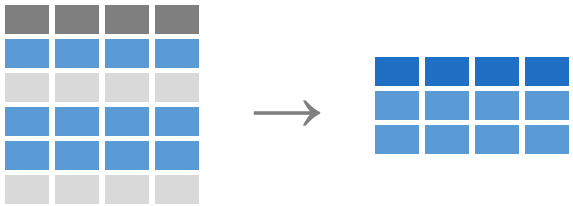
Population, Sample, Parameter, Statistic

- The **population** is every case you care about. Usually unseen.
- A **sample** is the subset you actually measured.
- A **parameter** describes the population. You want it, you rarely have it.
- A **statistic** describes the sample. You compute it, and it varies.

rep_slice_sample()

Take **repeated** random samples from a data frame. Each replicate is a fresh sample, tagged with its own replicate number.

Sample Again & Again



rep_slice_sample()

Example

the population to sample from

```
virtual_shovel <- bowl |>
  rep_slice_sample(n = 50)
```

the verb (function)

how many balls in one scoop

summarize()

Reduce each sample to a single number. With repeated samples there is no **group_by()** needed: the replicates are already grouped.

Example

```
virtual_shovel |>
  summarize(prop_red = mean(color == "red"))
```

the verb (function)

what to call the result

the proportion that are red

reps =

Ask for a thousand samples instead of one, and the statistic you get back becomes a **distribution**.

Example

```
virtual_samples <- bowl |>
  rep_slice_sample(n = 50, reps = 1000)

virtual_prop_red <- virtual_samples |>
  summarize(prop_red = mean(color == "red"))
```

1,000 separate samples

one proportion per replicate

sampling distribution

The distribution of a **statistic** across repeated samples. Not the distribution of the data: the distribution of the estimate.

How Much Does It Vary?

Its center sits near the parameter. Its spread is the standard error.

geom_histogram()

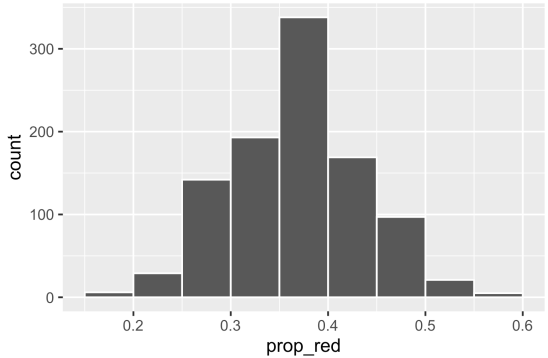
Look at the sampling distribution the same way you look at any other numerical variable.

the statistic, not the data

```
ggplot(virtual_prop_red, aes(x = prop_red)) +
  geom_histogram(binwidth = 0.05, boundary = 0.4, color = "white")
```

bin width

Example



Standard error

One number for how much an estimate bounces around.

The Spread Of The Estimate

- The **standard error** is the standard deviation of the sampling distribution.
- Bigger samples give a **narrower** sampling distribution.
- Quadruple the sample size to halve the standard error.
- The center barely moves; only the spread changes.

The Central Limit Theorem

Why so many sampling distributions look the same shape.

Bell-Shaped, Eventually

- As the sample size grows, the sampling distribution of the **mean** becomes bell-shaped.
- This holds even when the population itself is skewed.
- It is what lets a single sample say something about the population.
- The next chapter turns that into a **confidence interval**.