

Covariance, Correlation, and the Bootstrap

PSCI 1801 · Class Handout

Dr. Marc Trussler

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Covariance

We want a single number that describes how two continuous variables move together. Not “difference in means” — something that uses all levels of x .

$$\hat{\sigma}_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The average product of the deviations from each variable’s mean. Nothing is squared, so covariance can be positive or negative.

Intuition: Split the scatterplot into quadrants around $(\bar{x}, \bar{y})$:

- Points in the upper-right or lower-left → positive products → positive covariance
- Points in the upper-left or lower-right → negative products → negative covariance

R function: `cov(x, y)`.

Problem with covariance

It has no natural scale. Multiplying both variables by 100 multiplies the covariance by 10,000. That makes it hard to compare across measurements.

Correlation

Standardize the covariance by dividing by the two standard deviations:

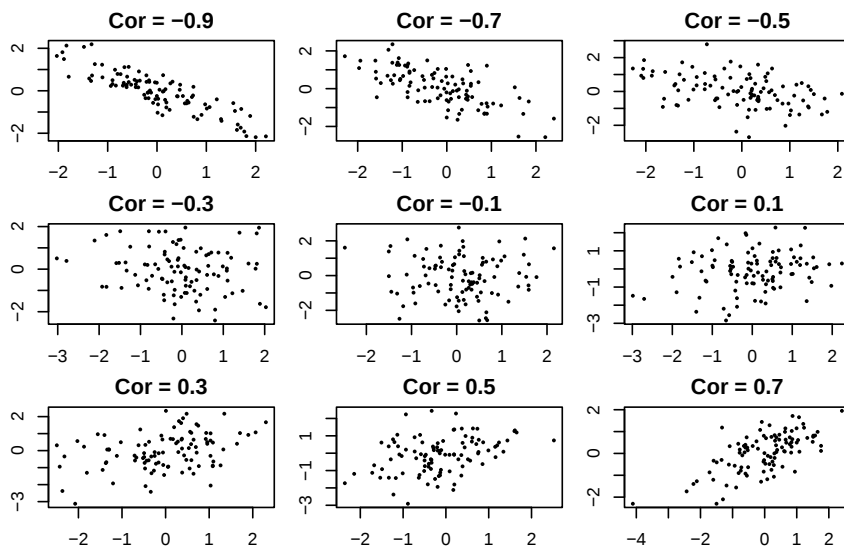
$$\hat{\rho}_{x,y} = \frac{\hat{\sigma}_{x,y}}{\sqrt{V[x] \cdot V[y]}}$$

Correlation is unitless and always between -1 and $+1$.

- $\rho = 1$: perfect positive linear relationship
- $\rho = 0$: no linear relationship
- $\rho = -1$: perfect negative linear relationship

R function: `cor(x, y)`. Use `use = "pairwise.complete"` if there are NAs.

A range of correlations, hard to eyeball:



The Bootstrap

The problem: `cor()` doesn't return a standard error. And there's no clean formula for the SE of a correlation.

The solution: the bootstrap — an all-purpose way to generate a sampling distribution for any statistic, using only the sample.

The idea

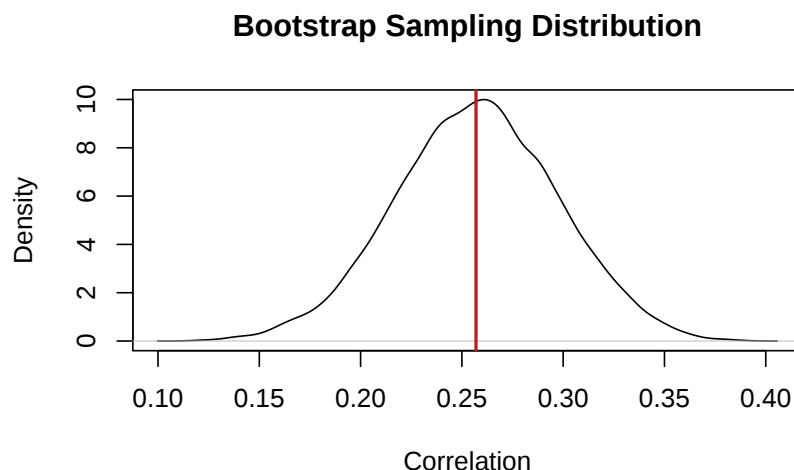
If your sample is representative of the population, then a resample of your sample (with replacement) is like a new sample from the population.

The procedure

1. Draw a new sample of size n from your original sample, **with replacement**
2. Compute the statistic of interest on this bootstrap sample
3. Repeat many times (e.g., 10,000)
4. The distribution of those bootstrap statistics approximates the sampling distribution

Sampling with replacement is essential — without it you just re-order your original sample and get the same statistic every time.

Example — bootstrapping a correlation



What to do with bootstrap output

Two things: a 95% confidence interval via quantiles, or a hypothesis test via the standard error.

Percentile CI:

```
      2.5%      97.5%  
0.1771611 0.3347370
```

Hypothesis test (assuming normality of the sampling distribution):

```
[1] 1.484659e-10
```

Key idea: the bootstrap distribution isn't centered on the true population value (we never learn that from one sample), but its width is an excellent estimate of the true standard error. That's all we need.