Beyond Means: Sampling Distributions of Other Common Statistics

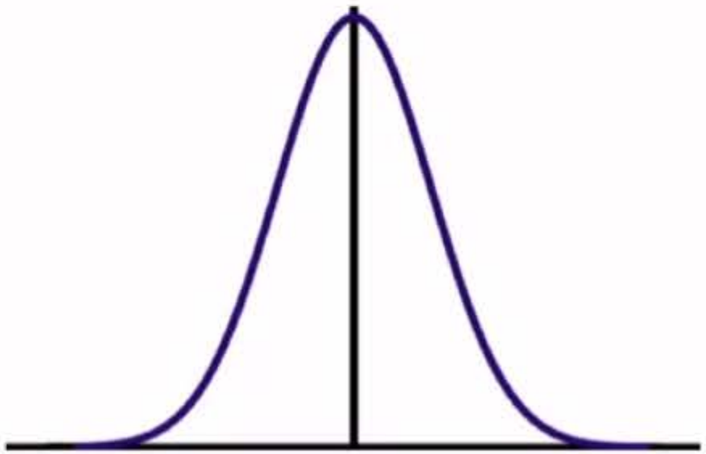
Brady T. West

Research Associate Professor, Survey Research Center, Institute for Social Research

An Interesting Result…

- Given large enough samples, **sampling distributions of most statistics of interest tend to normality** (regardless of how the input variables are distributed)

- This (**Central Limit Theorem**) result drives **design-based** statistical inference, or **frequentist** inference.

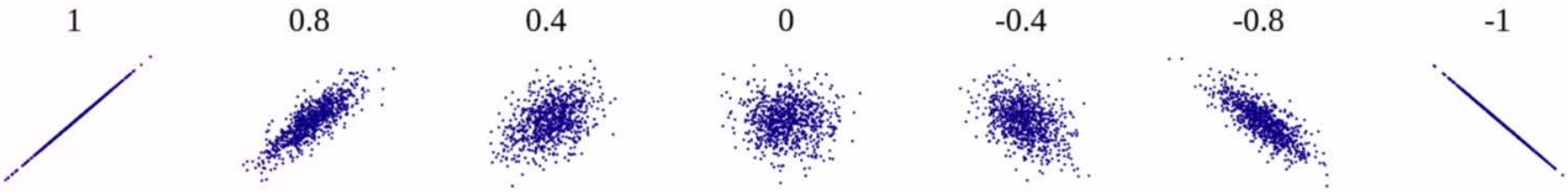


All possible values of the statistic

the distribution of estimates will tend to look like a bell shaped curve,

Simulation: Pearson Correlations

Consider Pearson's correlation coefficient, which describes the linear association between two continuous variables



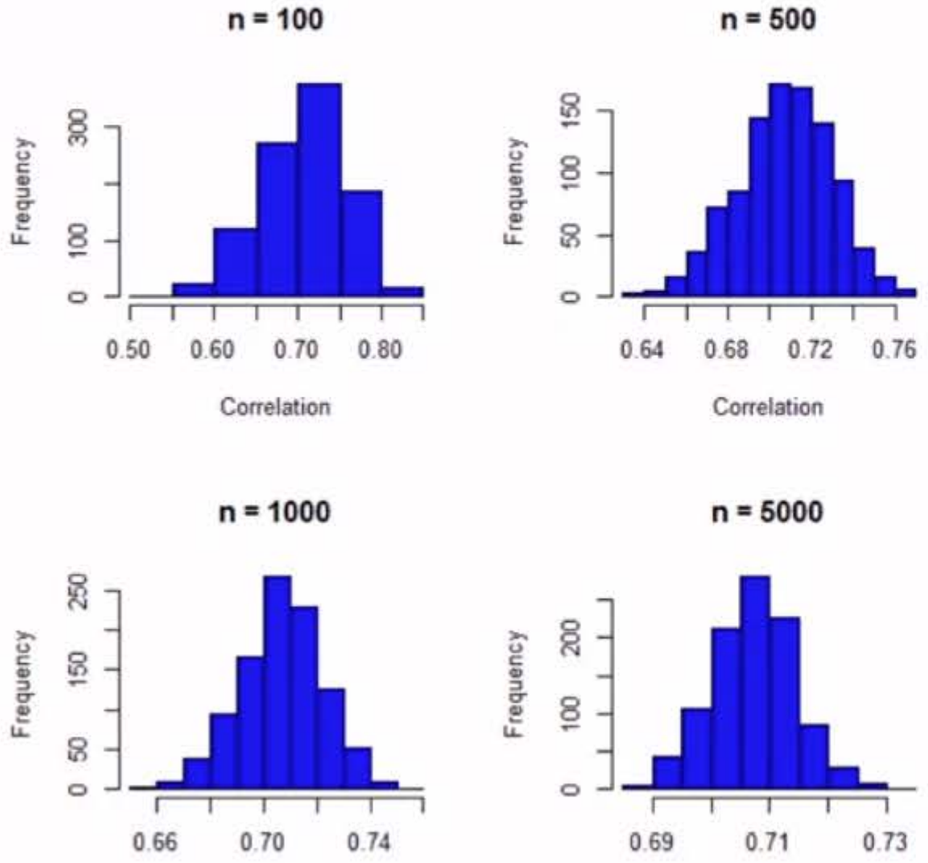
The correlation coefficient describes the linear association between two variables.

Simulation: Pearson Correlations

- **Simulate sampling distributions** for a correlation statistic:
  - Suppose **true population correlation is 0.7** (strong, positive)
  - Will **take 1,000 samples** of a specified sample size n
  - Do this for **various sample sizes** n = 100, 500, 1000, 5000

and then 1,000, and then 5,000,

Simulation: Pearson Correlations



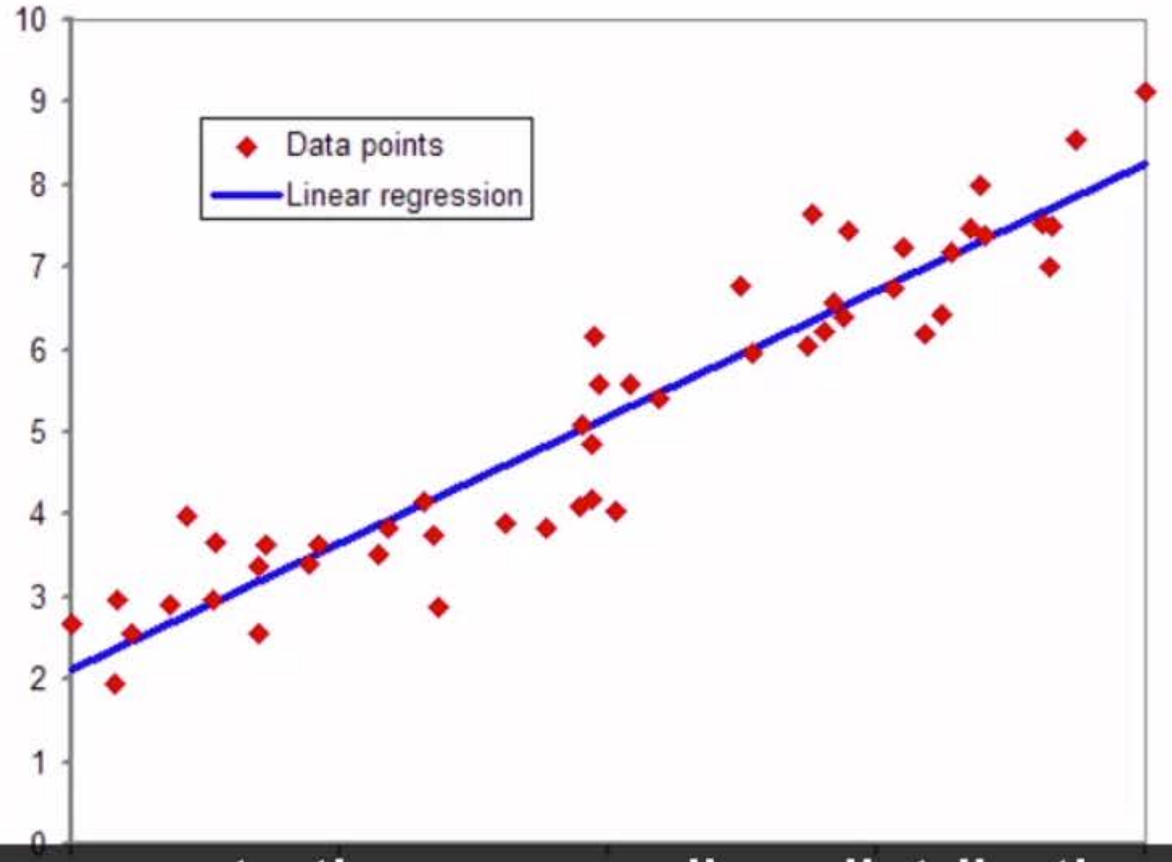
What do you notice about these sampling distributions?

- all approx. normal, centered at true correlation (0.7)
- as sample size $n \uparrow$ more symmetric and less spread

the population parameter is with a larger sample size.

Simulation: Regression Coefficients

Consider the **estimated slope**
(estimated change in y for a one unit ↑ in x)
for a **linear relationship**
between two continuous
variables



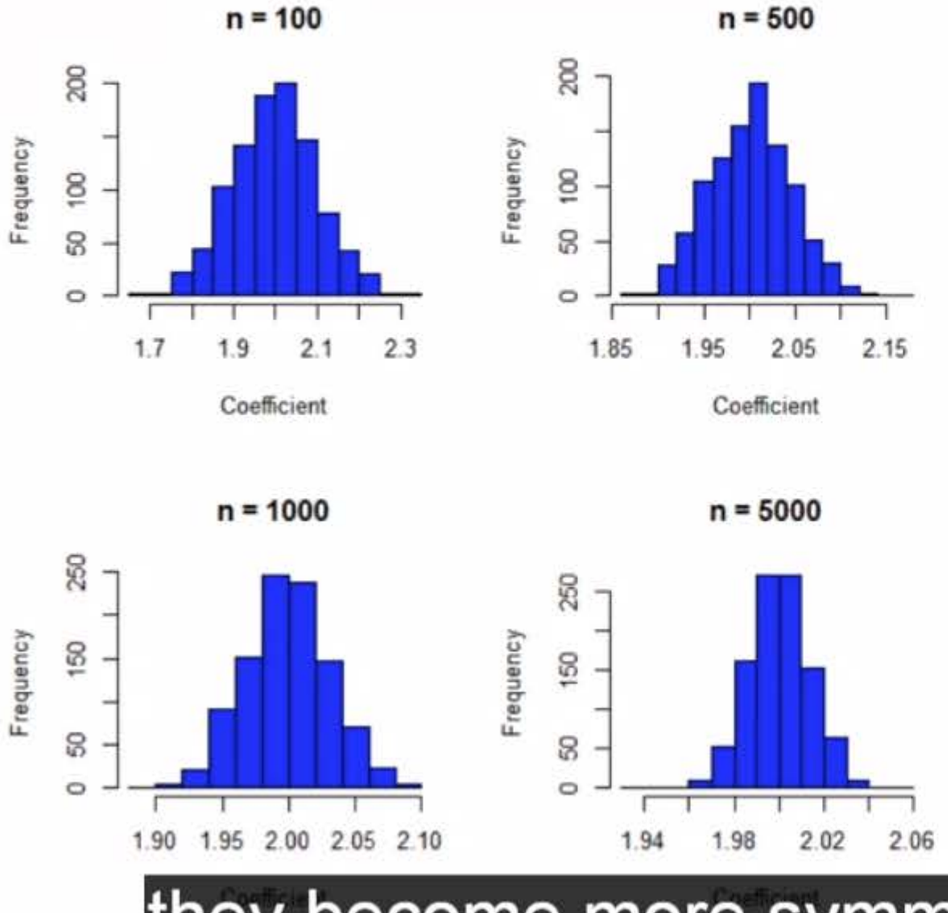
samples of different sizes to see what happens to those sampling distributions.

Simulation: Regression Coefficients

- **Simulate sampling distributions** for a **slope** statistic:
  - Suppose **true linear relationship in the population is** $y = 2x + \text{error}$, **so true slope is 2.**
  - Will **take 1,000 samples** of a specified sample size n
  - Do this for **various sample sizes** n = 100, 500, 1000, 5000

and we're going to do this for that same set of four different hypothetical sample sizes;

Simulation: Regression Coefficients



What do you notice about these sampling distributions?

- all approx. normal, centered at true slope (2)
- as sample size n ↑ more symmetric and less spread

they become more symmetric and there's less spread.

Sampling Distribution Properties

- Properties of sampling distributions for many popular statistics (regardless of complexity):
  - Normal, symmetrical, and centered at the true value
  - Larger sample sizes → less variability in estimates!

Key Point:

Can estimate variances of these normal distributions

based on only one sample

→ Enables **INFERENCE!**

the larger population while also accounting for sampling variability.

Suppose that the parameter you are interested in estimating for a given population is a proportion. That is, what fraction of individuals in a population has a characteristic of interest? This means that the variable of interest is a binary variable, taking on values of 1 (for those with the characteristic) and 0 (for those without the characteristic). If we select a large probability sample in order to estimate the proportion, what can we expect about the sampling distribution for this estimated proportion?

- The sampling distribution will not be normal, because the variable only takes on two possible values.
- The sampling distribution will be normal and centered at the true population proportion, and the larger the sample size, the more variance that the sampling distribution will have.
- The sampling distribution will be normal and centered at the true population proportion, and the larger the sample size, the less variance that the sampling distribution will have.

Correct

A proportion is the mean of a binary variable, and a commonly estimated parameter. The central limit theorem suggests that the sampling distribution of this type of mean will be approximately normal, with less variance as the sample size becomes larger. This is regardless of the fact that the variable of interest is binary.

- We don't have enough information to formulate an expectation.

Non-Normal Sampling Distributions

- **Not all** statistics have normal sampling distributions
- In these cases, **more specialized procedures needed** to make population inferences (e.g., **Bayesian methods**)

Cool example: variance components in multilevel models

(we will discuss these later in the specialization!)

and we'll talk about how to make inference in those cases.

What's Next?

So how exactly do we make population inferences based on one sample?

We can estimate features of the sampling distribution based on one sample…

but how do we get from that to population inference?