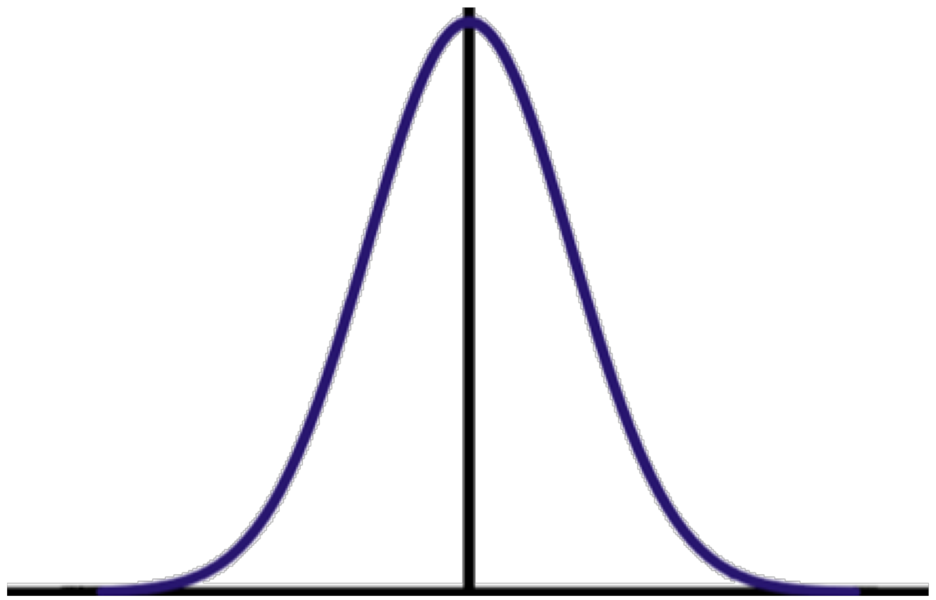
Beyond Means: Sampling Distributions of Other Common Statistics

Brady T. West

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)

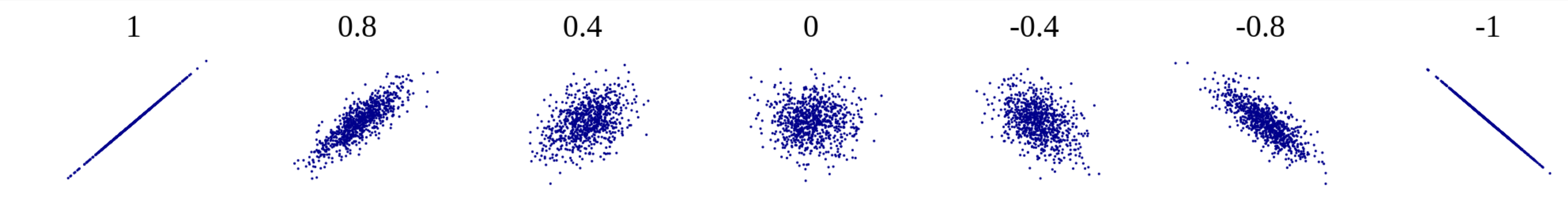
- This (**Central Limit Theorem**) result drives **design-based** statistical inference, or **frequentist** inference.



All possible values of the statistic

Simulation: Pearson Correlations

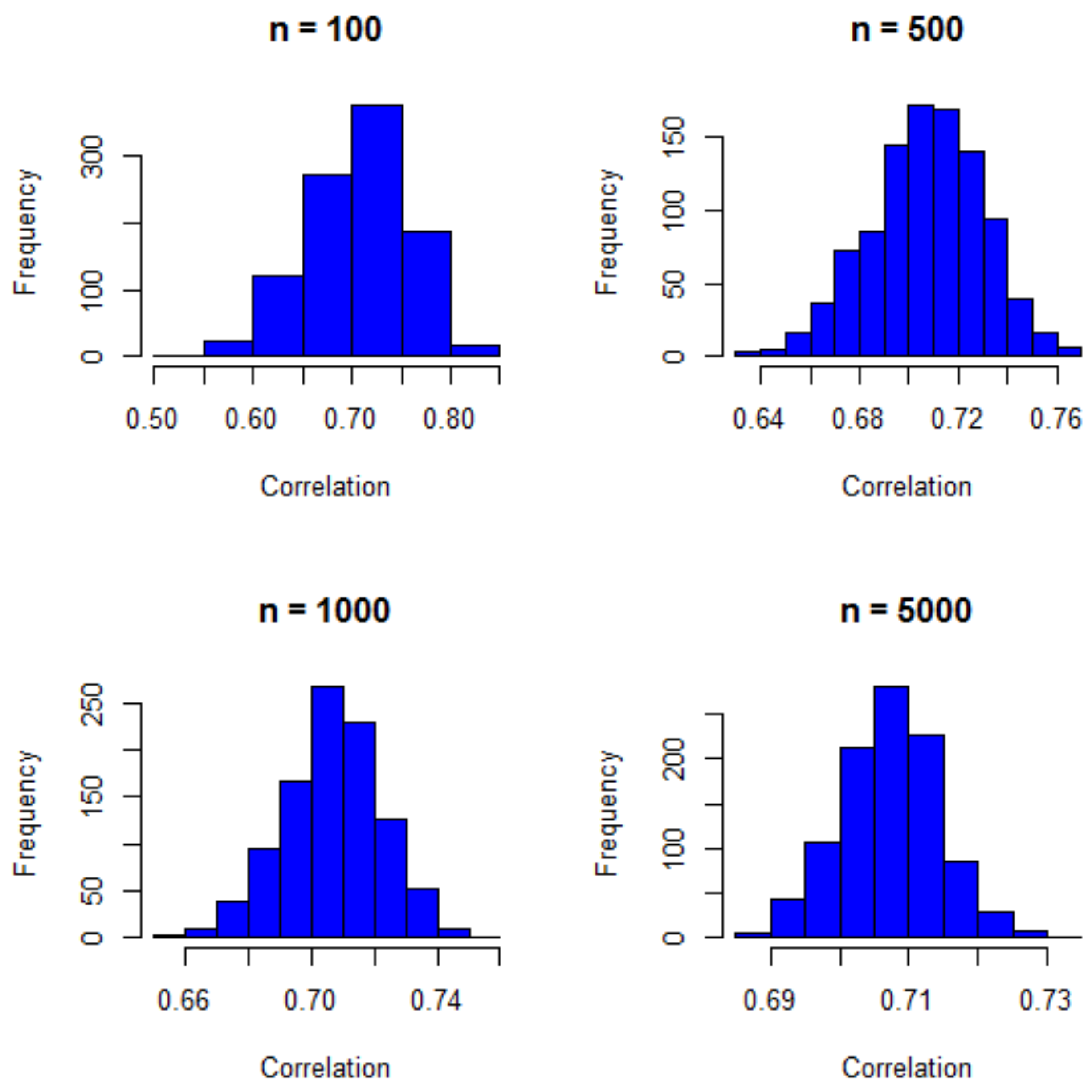
Consider Pearson's correlation coefficient, which describes the linear association between two continuous variables



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

Simulation: Pearson Correlations

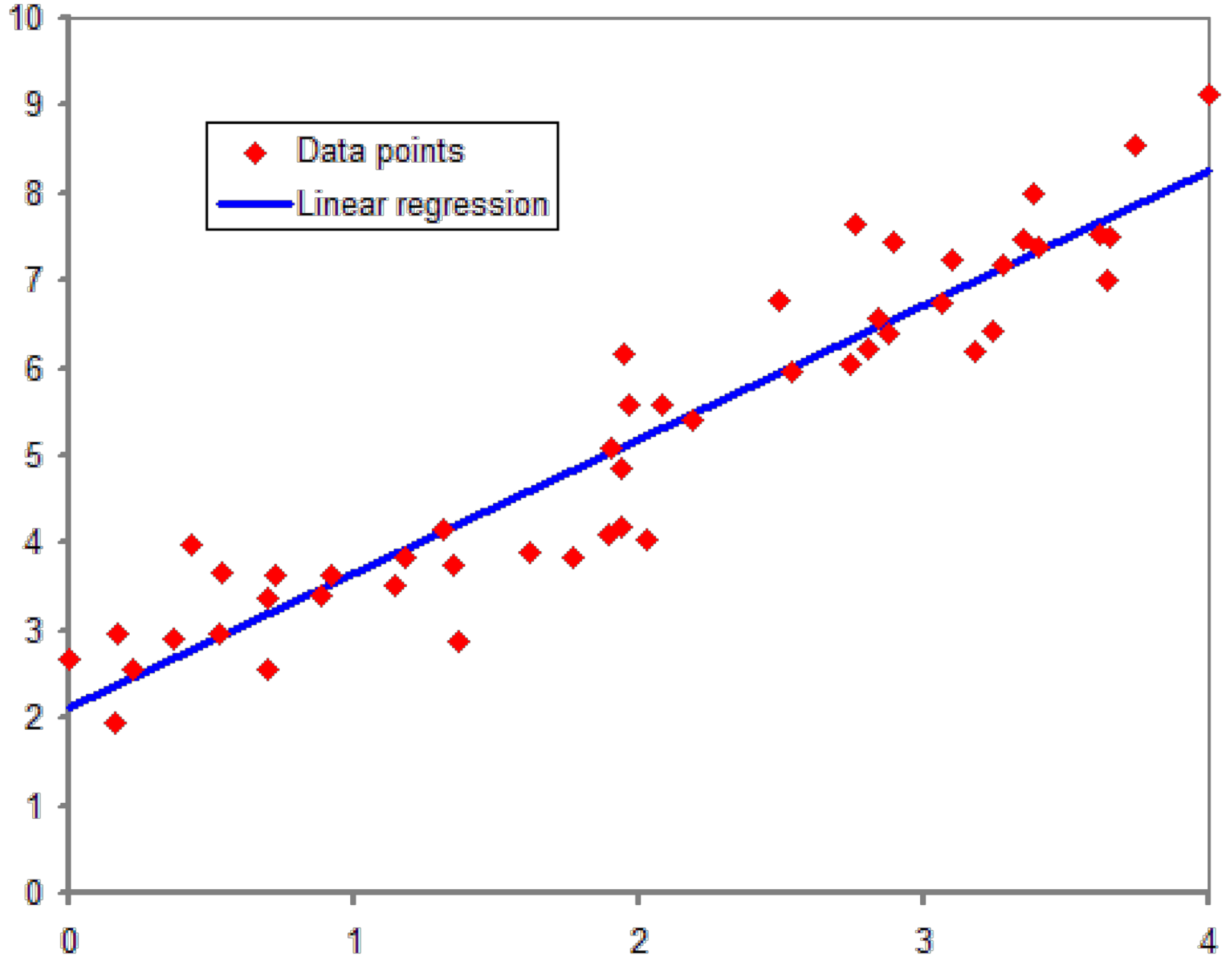


What do you notice about these sampling distributions?

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

Simulation: Regression Coefficients

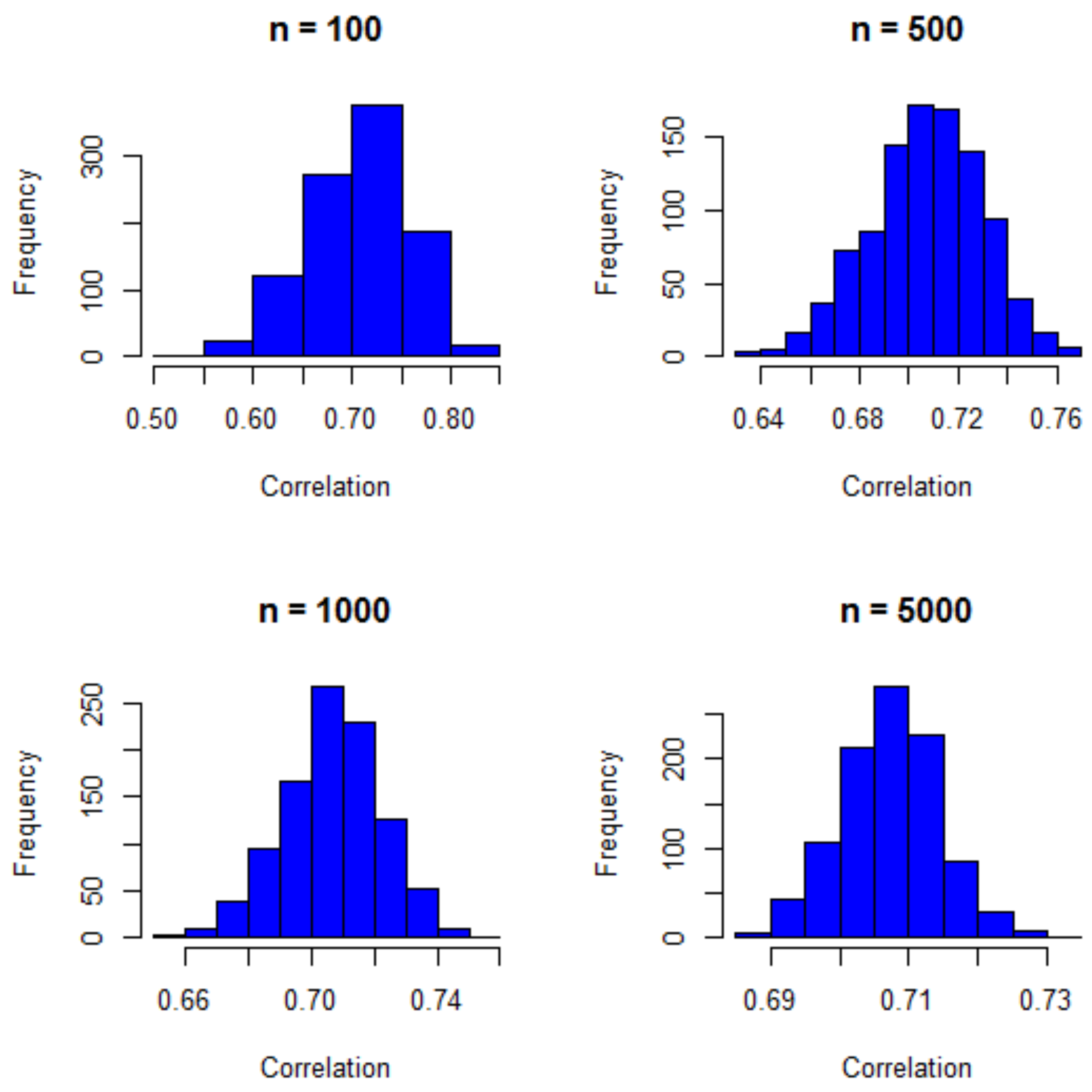
Consider the **estimated slope**
(estimated change in y for a one unit ↑ in x)
for a **linear relationship**
between two continuous
variables



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

Simulation: Regression Coefficients



What do you notice about these sampling distributions?

- all approx. normal, centered at true slope (2)
- as sample size $n \uparrow$ more symmetric and 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!**

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!)

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?