NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W4: 9/8-10/15

NC STATE UNIVERSITY

Administrative

- Appointments
  - Please send email in advance
- Moodle usage
  - Please be active, I will monitor and use that information for grading (as class participation)

NC STATE UNIVERSITY

Key points from 9/8

- Expected Values
  - The mean is the center of the distribution
  - The variance and standard deviation are measures of horizontal spread or dispersion of the random variable
- Basic Differential and Integral Calculus
- Maximum Likelihood Estimation
  - $L(\theta)=f(x_i; \theta_j)$, $l(\theta)$, $l'(\theta)=0$
  - Key properties

NC STATE UNIVERSITY

Today

- Central Limit Theorem
- Confidence Intervals
- Significance Testing

NC STATE UNIVERSITY

Always remember the differences

- Population
  - size: N (or ∞)
  - Parameters: mean (μ), variance (σ^2) $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$
- Sample
  - size: n (always finite)
  - statistic: mean ($\bar{x}$), variance (s^2) $s^2 = \frac{\sum(X-M)^2}{N-1}$
- Different samples (from same population) may produce different $\bar{x}$'s. The variation in these $\bar{x}$'s is called the standard error of the mean.

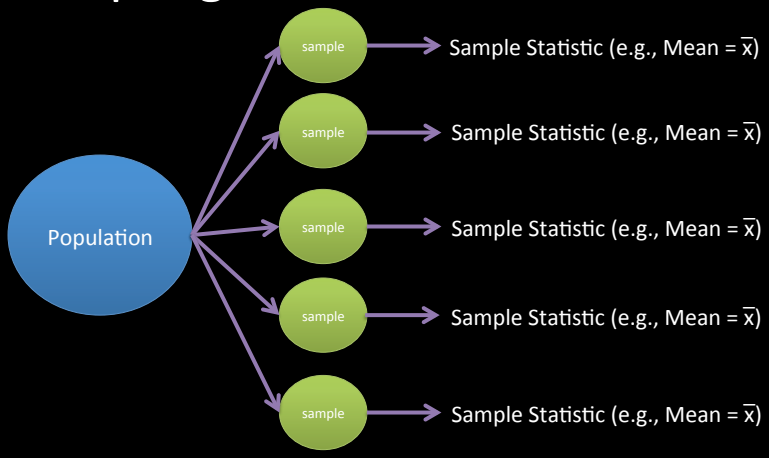
NC STATE UNIVERSITY

Bias of Estimator

- Bias: a statistic is biased if the mean of the sampling distribution of the statistic is not equal to the parameter it is estimating.
- Sampling Distributions: Suppose that we draw all possible samples of size n from a given population. Suppose further that we compute a statistic (e.g., a mean, median, standard deviation) for each sample. The probability distribution of this statistic is called a sampling distribution.

NC STATE UNIVERSITY

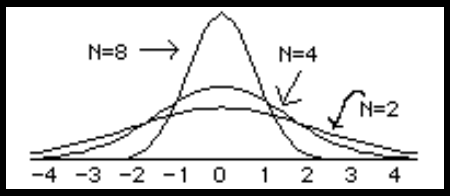
Sampling Distributions



NC STATE UNIVERSITY

Sampling distribution of the mean

- Sampling distribution is used in constructing CI for mean and for significance testing.
- Given a population with a mean of μ and a standard deviation of σ, the sampling distribution of the mean has a mean of μ and a variance of $\sigma_m^2 = \frac{\sigma^2}{N}$



NC STATE UNIVERSITY

Sampling Variability

- The sampling variability of a statistic refers to how much the statistic varies from sample to sample and is usually measured by its standard error; the smaller the standard error, the less the sampling variability.
- The standard deviation of the sampling distribution of the mean is called the standard error of the mean.

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

NC STATE UNIVERSITY

Central Limit Theorem

- Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases.

  $\bar{x}$ ~ N(mean = μ, and SE = s/√n)
- What is the significance?

  Regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases.

NC STATE UNIVERSITY

Conditions for CLT

1. Independence: sampled observations must be independent
   1. If sampling without replacement, n < 10% of population
2. Skewed population (or if we don't know the distribution)
   1. Sample size should be large (rule of thumb: n >30)
3. Note: According to CLT if population distribution is normal, sampling distribution is also normal, regardless of sample size

NC STATE UNIVERSITY

Example

- Let X be a random variable with μ = 10 and σ = 4. A sample of size 100 is taken from this population. Find the probability that the sample mean of these 100 observations is less than 9.
  - We can apply CLT : states that $\bar{x}$ follows approximately normal: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
  - To find probability (area under standard normal), first apply z transformation: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

NC STATE UNIVERSITY

Example

- We have:
  - μ = 10 and σ = 4, n = 100

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- $P(\bar{x} < 9) = P(z < (9-100) \div (4/\sqrt{100})) = P(z < -2.5)$
- Check value from standard normal probabilities table
- $P(z < -2.5) = 0.0062$

NC STATE UNIVERSITY

Confidence Intervals

- Properties of estimators provide valuable information for comparing and choosing an estimator, but no quality information
  - E.g., consistency guarantees that the estimated value will approach the true value in the limit, but does not give information on how close it is to the true value, given some data with a certain number of samples.
- For quantifying the quality of a particular estimate, we can compute a *confidence interval* (*CI*) around the estimate $\hat{\theta}_n$, such that this interval covers the true value θ with some high probability 1 − α. The narrower this interval, the closer we are to the true value, with high probability.

NC STATE UNIVERSITY

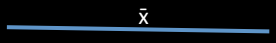
Example

- Suppose if I take a sample of 6 students from the class, and found that the mean is 150lb.
- The sample mean is the point estimate of population mean
- Point estimate is limited by itself, as it does not reveal the uncertainty associated with the estimate
- That is, we do not have a good sense of how far this sample mean may be from the population mean
- In other words, are we confident that the population mean is within 5 pounds of 150?

NC STATE UNIVERSITY

CI

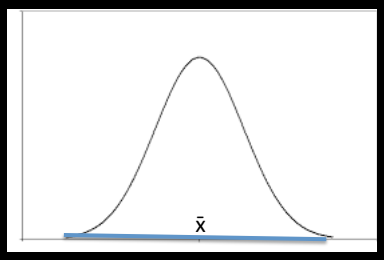
- CI: possible range of values for the population parameter
- Given a sample, $\bar{x}$ is our best guess
- So the range of values should be centered around our best guess



NC STATE UNIVERSITY

CLT

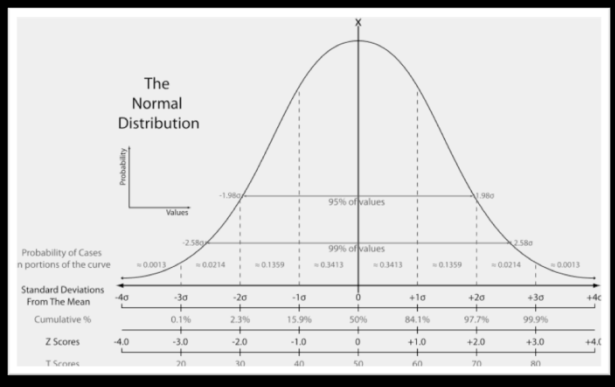
- Given a sample, $\bar{x}$ is our best guess
- So the range of values will be around our best guess
- $\bar{x}$ ~ N(mean = μ, and SE = s/√n)



NC STATE UNIVERSITY

CI

- Normal distribution



~95% CI = $\bar{x}$ ± 2SE

NC STATE UNIVERSITY

CI

- CI of population mean: Computed as the sample mean ± a margin of error
  - This is the critical value corresponding to the middle XY% of the normal distribution times the standard error of the sampling distribution

$$\bar{x} \pm z^* (s)/\sqrt{n}$$

NC STATE UNIVERSITY

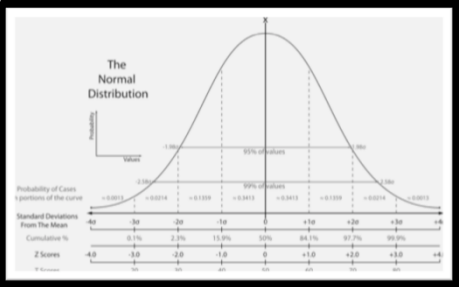
Condition for CI

- Similar to CLT
  - Independence
  - n ≥ 30 for skewed population distributions

NC STATE UNIVERSITY

How to find critical value

- $\bar{x} \pm z^* (s)/\sqrt{n}$
- We can use table
- For 95% CI, we have (1-0.95)/2 = 0.025
- From table z* = 1.96



NC STATE UNIVERSITY

General Outline for Finding CI

- (1) Identify sample statistic (e.g., mean)
- (2) Select a confidence level (e.g., 90%, 95%, …)
- (3) Compute margin of error:
  - Critical value * Standard deviation of statistic

  OR
  - Critical value * Standard error of statistic
- (3.1) Computing Critical Value:
  - The central limit theorem states that the sampling distribution of a statistic will be nearly normal
  - When the sampling distribution is nearly normal, the critical value can be expressed as a z score.

NC STATE UNIVERSITY

General Outline for Finding CI

- (3.1) Computing Critical Value (continued):
  - When the sampling distribution is nearly normal, the critical value can be expressed as a z score.
    - First compute following quantities:
      - $\alpha = 1 - (\text{confidence level} / 100)$
      - critical probability (p^*): $p^* = 1 - \alpha/2$
    - To express critical value as z score: find the z score having a cumulative probability equal to the critical probability (p^*)
- (4) Finally, CI = Sample statistic ± Margin of error

NC STATE UNIVERSITY

Example

- A sample of 30 students is drawn from CSC-522 population of 100. Average weight of sample is 150 pounds and standard deviation is 20 pounds. Compute the 95% CI.
- Use the 4-step solution outlined in previous slides to find the answer