PRELIM UNDERSTANDING – POPULATION vs SAMPLE

- We want to understand the relationship between two variables in the population but we do not have data for every person in the population.
- Take the data for a smaller sample drawn from the population.
- If the sample is "large enough" and drawn randomly from the population, then we can make inferences about the population from the relationships observed in the sample.
- The reason we can draw inferences is because of two fundamental theorems in probability:
  - "Law of Large Numbers"
  - "Central Limit Theorem"

PRELIM UNDERSTANDING

- Suppose that we draw all possible samples of size n from a given population.
- We compute a statistic (e.g., a mean, proportion, standard deviation) for each sample.
- The probability distribution of statistic is called a sampling distribution (say Y-dash).
- Now, as n increases,
  - Mean of sample (or Y-dash) becomes more tightly centered around mean of population.
  - Distribution tends to become more normal.

PRELIMS - STATISTICAL SIGNIFICANCE

How much confidence can we have in the values of β_0 and β_1 estimated from our first sample? (what if another sample provide slightly different values.)

We need to test the hypothesis that there is indeed a non-zero correlation between Y and X which translates to testing the null hypothesis: $\beta_0 = \beta_1 = 0$

HYPOTHESIS TESTING

Test inferences about population parameters using data from a sample.

In order to test a hypothesis in statistics, we must perform following steps:

1. Formulate a null hypothesis and an alternative hypothesis on population parameters.

$$H_0: \overline{X} = \mu \quad \text{vs. } H_A: \overline{X} \neq \mu$$

2. Build a statistic to test the hypothesis made.

$$z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

where $\overline{X}$ is the sample mean and s is the sample standard deviation.

3. Define a decision rule to reject or not to reject the null hypothesis.

Two Tail Hypothesis Test

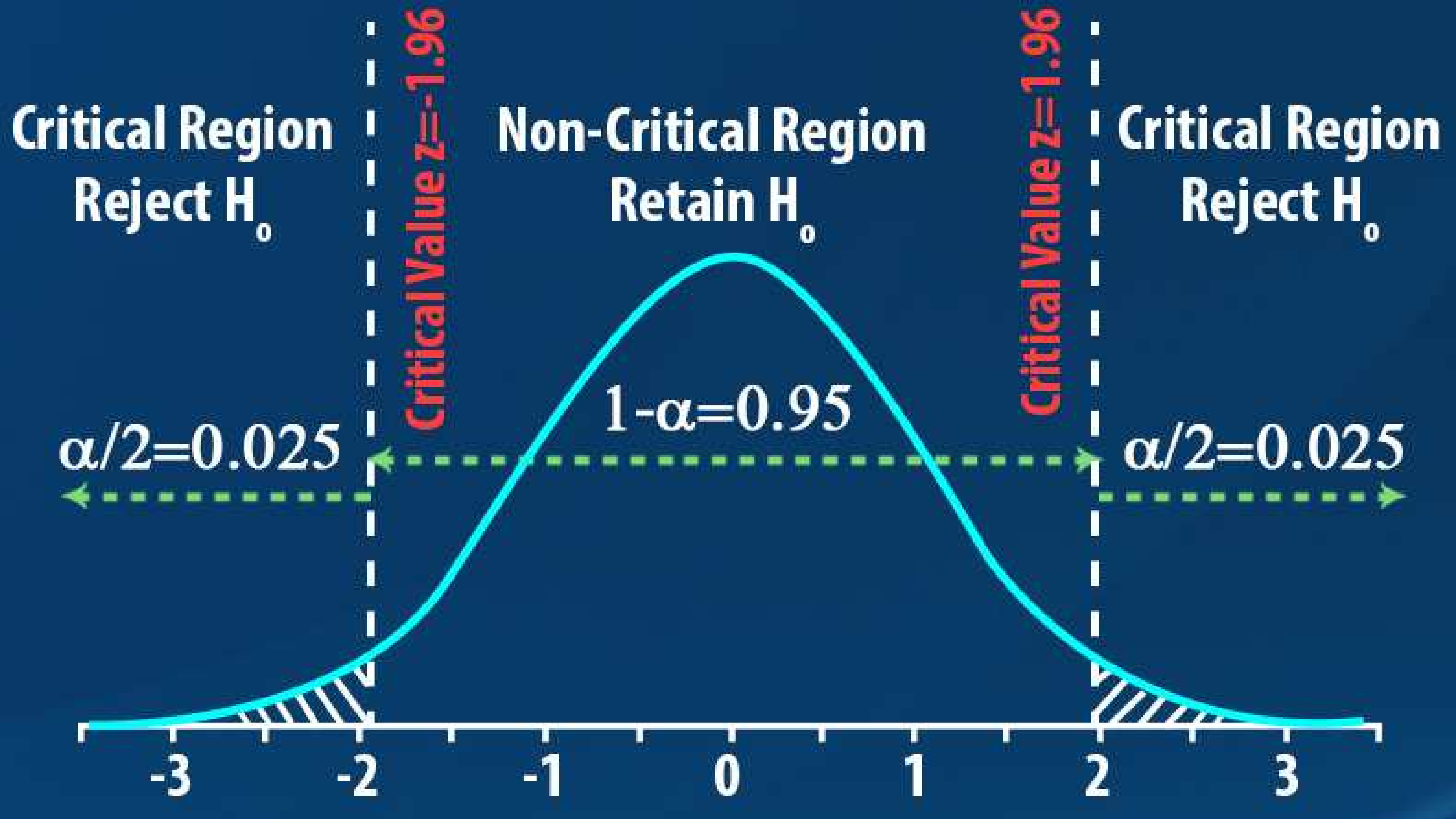

Figure:2

STATISTICAL SIGNIFICANCE – P VALUES

- We have to test the null hypothesis estimated $\hat{\beta}_1$=0, that is there is no correlation.
- We find the standard error associated with the estimated betas as follows:

$$SE(\hat{\beta}_1) = \frac{\sqrt{\Sigma(Y_i - \hat{Y}_i)^2}}{(n-2)\sqrt{\Sigma(X_i - \bar{X})^2}}$$

  - Given the estimate of $\hat{\beta}_1$ and the standard error of the estimate $= SE(\hat{\beta}_1)$
  - We calculate a t-statistic for $\hat{\beta}_1$: $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$
- If t-statistic > 1.96, we can reject null hypothesis $\hat{\beta}_1 = 0$ with 95% confidence.
- The p-value associated with each variable gives the probability that we could have observed the value of $\hat{\beta}_1$ or larger, if the true value of β was in fact 0.
- Very small p-values indicate there is a very small probability of the real β being 0.
- Indicates that there is a statistically significant relationship between Y and X that is not just due to chance alone.